The Cambridge Handbook of

English Corpus Linguistics

edited by Douglas Biber and Randi Reppen

The Cambridge Handbook of English Corpus Linguistics

Edited by
Douglas Biber
and
Randi Reppen
Northern Arizona University

CAMBRIDGE
UNIVERSITY PRESS

5

Keywords

Jonathan Culpeper and Jane Demmen

1 Previous research on keywords

1.1 An introductory survey

The term "keyword" has considerable currency outside corpus linguistics, but there it is usually understood in a different sense. Raymond Williams's ([1976] 1983) book, entitled *Keywords*, was not the first to discuss keywords (see, for example, Matoré 1953). However, it has done more than any other publication to popularize the idea that some words – such as *democracy, private*, or *industry* – are "key" because they capture the essence of particular social, cultural, or political themes, thoughts or discourses. In fact, Williams's work has given rise to a journal: *Keywords: A Journal of Cultural Materialism*. Such words are deemed key on the basis of "readings" of their role in representing and shaping culturally important discourses. In corpus linguistics, a keyword has a quantitative basis: it is a term for a word that is statistically characteristic of a text or set of texts. The advantages of this basis are twofold. First, it is less subject to the vagaries of subjective judgments of cultural importance. Second, it does not rely on researchers selecting items that might be important and then establishing their importance, but can reveal items that researchers did not know to be important in the first place.

However, we would not want to argue that the two different kinds of keyword, qualitatively defined and quantitatively defined, are completely separate. Quantitatively defined keywords can indeed be of social, cultural, or political significance if they are characteristic of social, cultural, or political texts. That significance can be established by conducting qualitative analyses of the keywords identified through initial quantitative analysis. This has been the tactic of the bulk of studies applying the quantitative notion of keywords to particular texts, in a wide range of genres. These include, for example, Baker (2004) on gay and lesbian texts and Baker (2009b) on parliamentary debates, Fairclough (2000) and Jeffries and Walker (2012) on newspaper representations of New Labour, Johnson

et al. (2003) on newspaper political correctness discourse, McEnery (2009) on books by social activist and campaigner Mary Whitehouse, Gerbig (2010) on travel writing, and Warren (2010) on engineering texts. Studies of literary texts include Archer and Bousfield (2010) and Culpeper (2002) on characters in Shakespeare's plays, Fischer-Starcke (2009) on Jane Austen's novels, Tribble (2000) on romantic fiction and, in the area of present-day drama, Bednarek (2010) on TV situation comedy and McIntyre (2010) on film dialogue. Research into languages other than English includes Fraysse-Kim (2010) on Korean school textbooks and Philip (2010) on Italian political speeches and press releases.

An important strand of work has focused on the nature of keywords (henceforth, the term *keyword(s)* will always refer to quantitatively defined items, unless specified otherwise). Discussions of keywords can most clearly be traced back to work in stylistics, particularly statistical stylistics, one of the early pioneers being Guiraud ([1954] 1975), but also to frequency-oriented discussions of style, notably that of Nils Erik Enkvist (e.g. 1964) on "style markers" (see Section 1.2). However, it was not until the advent of computer programs, and especially Mike Scott's *WordSmith Tools* (1996–2013), that we saw an explosion in studies, such as those listed in the previous paragraph. Notable research includes Scott's own, culminating in especially Scott and Tribble (2006) (see Section 1.3). By the mid 2000s, the field of keyword study had reached a certain maturity, with scholars having grappled with the notion of a keyword, applied it in analyses of discourses, and tackled methodological problems. Baker (2004) is a good example of a study that both performs a keyword analysis of a particular set of texts and reflects on the nature of that analysis. In the area of stylistics, Stubbs (2005) demonstrates the advantages of extending the analysis of keyness from single words to multi-word units (see Section 1.4). Following some reflections on landmark studies in keyness, we provide a more focused discussion of keyness methodology, in Section 1.5. At the end of Section 1, we sum up the state of the art in keyness research (in Section 1.6), and in Section 1.7 we highlight two relatively new approaches to investigating key items other than single words: key parts of speech (POS) and key semantic domains. In Section 2 we present a case study of the application of key POS and key semantic domains from Culpeper (2009b), which evaluates the benefits they offer in addition to an analysis of keyness at the word level.

1.2 Early history

Some notion that relatively frequent words can characterize particular literary styles, notably authorial styles, has been around for centuries, especially in French stylistics (Ullmann 1973: 72–73 cites authors from as far back as 1832). But it is studies in the area of statistical/computational stylistics or stylometry that present the clearest line of descent to the notion of a keyword used in corpus linguistics today. Perhaps the first to

use the term "keyword" ("mot-clés") for this particular concept was Pierre Guiraud ([1954] 1975). Guiraud (1975: 64–66) contrasts "mots-clés" (based on relative frequency) with "mots-thèmes" (based on absolute frequency):

> Toute différente est la notion de *mots-clés*, qui ne sont plus considérés dans leur fréquence absolue, mais dans leur fréquence relative; ce sont les mots dont la fréquence s'écarte de la normale.
>
> [Wholly different is the notion of *mots-clés* (keywords), which are not considered in terms of their absolute frequency, but their relative frequency; these are the words whose frequency diverges from the normal.]

Simply being relatively statistically significant is not in itself the important point of interest. That lies in the link between keywords and style, but it is not articulated by Guiraud. Although he does not use the label "keywords," this link is clearly articulated by Nils Erik Enkvist (1964). In the following quotations, Enkvist defines style in terms of "frequencies," "probabilities," and "norms," and goes on to define "style marker":

> Style is concerned with frequencies of linguistic items in a given context, and thus with *contextual* probabilities. To measure the style of a passage, the frequencies of its linguistic items of different levels must be compared with the corresponding features in another text or corpus which is regarded as a norm and which has a definite relationship with this passage. For the stylistic analysis of one of Pope's poems, for instance, norms with varying contextual relationships include English eighteenth-century poetry, the corpus of Pope's work, all poems written in English in rhymed pentameter couplets, or, for greater contrast as well as comparison, the poetry of Wordsworth. Contextually distant norms would be, e.g., Gray's *Anatomy* or the London Telephone Directory of 1960. (1964: 29, original emphasis)
>
> We may . . . define style markers as those linguistic items that only appear, or are most or least frequent in, one group of contexts. In other words, style markers are contextually bound linguistic elements . . . style markers are mutually exclusive with other items which only appear in different contexts, or with zero; or have frequencies markedly different from those of such items. (1964: 34–35)

Enkvist's concept of style markers as words whose frequencies differ significantly from their frequencies in a norm corresponds precisely to what keywords are. Repetition is the notion underlying both style markers and hence keywords, but not all repetition, only repetition that statistically deviates from the pattern formed by that item in another context.

1.3 Establishment in corpus linguistics

It is in the context of corpus linguistics that the notion of keywords and the practice of keyword analysis has been developed and popularized, notably by Mike Scott through the KeyWords facility of WordSmith Tools (Scott

1996–2013), a program making it a relatively easy and rapid task for a researcher to calculate the incidences of each and every single word in the target data as well as a comparative dataset, undertake statistical comparisons between incidences of the same words in order to establish significant differences, and see the resulting keywords ranked according to degrees of significance of difference.

Scott and Tribble (2006: chapter 4) call attention to some of the most crucial issues in keyword analysis. They emphasize the nature of the comparative textual data, specifically the choice and type of "reference corpus" (see Section 1.5). They also argue that (apart from proper nouns, which often tend to appear), keywords tend to be of two main types: those relating to the text's "aboutness" or content, and those which are related to style (see further Scott 2013: 196). Scott (2000: 155) links aboutness to Halliday's (e.g. 1994) ideational metafunction, and also suggests that aboutness keywords are those that we would be "likely to predict" (Scott 2000: 160). Scott and Tribble's (2006: 60) analysis of Shakespeare's play *Romeo and Juliet* reveals the following as such keywords: *love*, *lips*, *light*, *night*, *banished*, *death*, and *poison*. But their analysis also shows that some items – exclamations, *thou*, *art*, and the pronoun *she* – do not fit the notion of aboutness, for which they propose the label "style" as a cover term (ibid.) (see Culpeper 2009b: 39 for an alternative categorization).

Scott and Tribble (2006) also illustrate the dispersion of keywords (see further Section 1.5). They examine the distribution of keywords in *Romeo and Juliet*, partly with the aim of demonstrating a "dispersion plot" (2006: 65–70). Figure 5.1 is generated by WordSmith Tools. Vertically, it displays

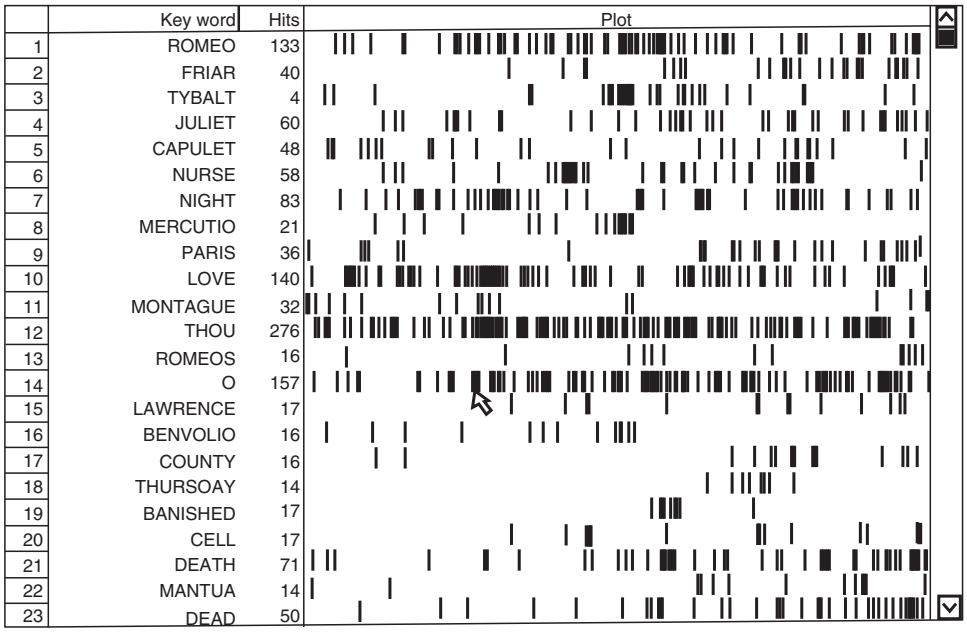

Figure 5.1 A keyword dispersion plot of *Romeo and Juliet* generated by *WordSmith Tools* (reproduced from Scott and Tribble 2006: 65)

keywords from the play in order of keyness; horizontally, the plot shows a small virtual line for every instance of the keyword, ordered from the beginning of the play on the left to the end of the play on the right.

The utility of this tool is obvious: a visual map of keyness enables us to see whether or not the keywords are likely to be general features of the play or concentrated at particular points. Scott and Tribble (2006: 66) make a distinction between "global" keywords, such as *thou*, which appears to be more or less evenly spread throughout the play, and "local" keywords, such as *banished*, which is mostly confined to a narrow area (Act III Scene ii, where the impact of Romeo's banishment is felt).

Connected with the issue of dispersion, Scott and Tribble (2006: 66–70) explore the idea that some keywords are not only key within a particular text but share collocational space with other keywords. For example, other keywords that collocate with the keyword *Romeo* within a 10-word collocational span (5 words to the left and 5 words to the right) are: *Benvolio*, *Juliet*, *banished*, *night*, *Tybalt*, *art*, *dead*, *O*, *thou*, and *is*. This suggests that there is a "natural association" (2006: 68) amongst the keywords. Thus, Romeo is friends with Benvolio, a lover of Juliet, banished by Capulet, meets at night, kills Tybalt, and so on. Scott and Tribble (2006: chapter 5) go on to explore linkages between keywords, in more detail, showing how these can be used to profile genres. They use the *British National Corpus* (BNC; see Aston and Burnard 1998) to investigate "key keywords" (keywords that are key in many texts) and "associates" (the co-key items associated with key keywords found in a number of texts) (see also Scott 1997). It should be noted that these notions can only be explored with a very large number of texts. Scott and Tribble (2006: 70–72) also analyze 1,000 randomly selected BNC texts in order to consider the POS characteristics of keywords. They report that four word classes (nouns, determiners, prepositions, and pronouns) account for 57 percent of keyword types, and that interjections, pronouns, alphabetical symbols, proper nouns, possessive *-s*, the verb *be*, and nouns are the most likely sources of keywords. They leave it for further research to investigate the reasons for this intriguing finding.

1.4 Further keyword developments

In his (2004) analysis of gay and lesbian texts, Baker discusses some of the major pitfalls in interpreting keywords. He shows the risks of overgeneralizing or overstating what keyness implies, emphasizing that "keywords only focus on lexical differences, rather than semantic, grammatical, or functional differences" (2004: 354). Baker not only points out that keywords may be unevenly dispersed in a corpus, and accounted for by only a minority of texts, but that keyness may involve only one of several possible senses of a word. The word *session*, for example, has a particular sexual meaning in the gay texts, which accounts for its keyness when compared to the lesbian texts, but if instances with this meaning are excluded then

session would not be a keyword (2004: 354–55). Baker airs the pros and cons of lemmatizing or annotating data, adjusting the settings of corpus tools to constrain the numbers of results generated, and looking at multi-word units (see below and Section 1.5). He notes that a problem with moving beyond the word level to, for example, semantic categories is that it involves making subjective choices about the categories (2004: 353).

As a way of finding keywords which are not concentrated in only a minority of texts, Baker (2004: 350–351) also investigates key keywords. He argues that key keyword results include "keywords that we could have probably made a good educated guess at in advance," for example that the lesbian texts contains "more female pronouns and more words relating to female parts of the body and clothing" than the gay texts. Baker considers this a limitation of key keywords (2004: 350), though it is worth pointing out that empirical evidence which supports intuitions about corpora can be useful, particularly if one is not very familiar with the data or text-type. Baker (2004: 349) also makes the important point that keywords are oriented to differences between texts, and argues that similarities should not be overlooked. Despite this, most studies utilizing keyness have not sought to contextualize differences by looking at empirically-based language similarities. Baker's (2011) more recent concept of "lockwords" offers a possible way forward (see also the brief comments in Section 1.6).

Other scholars have focused specifically on the benefits of looking at key multi-word units in addition to single words. Various terms are adopted for multi-word linguistic units, such as "lexical bundles" (Biber *et al.* 1999), "word clusters" (the items generated by Scott's program WordSmith Tools) and "n-grams," sometimes under a general heading of "phraseology." As might be expected, there are issues of compatibility, depending on the program and/or parameters used. Multi-word units are often structural fragments rather than complete units. They are best considered as recurring word sequences which have collocational relationships. Investigating key multi-word units has been particularly fruitful in the area of stylistics, and popularised by Stubbs' (2005) analysis of Joseph Conrad's novel *Heart of Darkness*. Stubbs argues that looking at single (key)words offers limited prospects for analysis, pointing out that words occur in recurrent "lexico-grammatical patterns" (2005: 15–18) and that "collocations create connotations" (2005: 14). Stubbs illustrates that recurrent multi-word units provide evidence of recurrent metaphors and/or intertextuality (2005: 13–14) which may not be discernible from single-word results, and that they can reveal pragmatic and discoursal aspects of dialogue and narrative. Other linguists with an interest in literary texts who have made notable contributions to key multi-word unit research include Mahlberg (e.g. 2007), who shows how some of the memorable characters in Charles Dickens' novels are created through recurrent localized word clusters, and Fischer-Starcke (e.g. 2009), who investigates how frequent phrases in Jane Austen's novels contribute to the construction of characters, and places such as the city of Bath.

1.5 Methodology

To generate a list of keywords, corpus linguistic software programs such as WordSmith Tools, AntConc, and WMatrix conduct statistical comparisons between the words in one dataset (also called the "target" corpus) and another (the comparative or "reference" corpus). This is done by first generating a "word list" of the lexical items in the dataset and the reference corpus, using the appropriate program function. The keyness tool is then run, which makes a series of frequency-based comparisons using a statistical significance test, typically Dunning's (1993) log-likelihood test or the chi-square test. The tool ranks the results in a list showing the keyness value (statistical significance) of each word. Positive keywords are those with comparatively high frequency, and negative keywords are those with comparatively low frequency (Scott e.g. 2013: 196).

The ease and rapidity of keyness analysis afforded by the programs mentioned above is the major boon. This boon is also apparent when one considers other methods designed to reveal styles. Using data comprising three genres (conversation, monologic speech, and academic prose), Xiao and McEnery (2005) compare a multidimensional analysis of the type conducted in Biber (1988) with a keyword analysis. The results they obtain are "similar" for both methods (Xiao and McEnery 2005: 76), but the keyword analysis was "less demanding" (2005: 77), for the reason that multidimensional analysis first involves some relatively complex algorithms for the extraction of certain grammatical features from the corpus, and then further relatively complex statistical analysis. However, the downside of the comparative ease of keyword analysis is the potential for performing studies in a relatively mechanical way without a sufficiently critical awareness of what is being revealed or how it is being revealed. Below, we will briefly dwell on some important considerations for the analyst seeking to generate maximally useful keyword results.

1.5.1 The reference corpus

Scott and Tribble (2006) conclude that "while the choice of reference corpus is important, above a certain size, the procedure throws up a robust core of KWs [keywords] whichever the reference corpus used" (2006: 64). Of course, a set of data which has no relationship whatsoever with the data to be examined is unlikely to reveal interesting results regarding the characteristics of that data (as Enkvist (1964) notes with regard to comparing a poem by Pope with a telephone directory; see Section 1.2 above). What if one selects a huge multi-genre corpus, such as the BNC, as indeed other studies have done (e.g. Johnson *et al.* 2003; Tribble 2000)? We can readily hypothesize that some genres within that corpus have a relatively close relationship with the data to be examined, whilst other genres have a relatively distant relationship. The choice of the reference corpus will affect the potential for acquiring keyword results that are relevant to the particular aspect of the text(s) one is researching (see also Scott 2013: 197).

For example, Johnson *et al.* (2003) are interested in what characterizes the discourse in which political correctness expressions appear in different newspapers. Some of the keywords they retrieve simply reflect the fact that it is newspaper discourse, and have nothing specifically to do with political correctness.

Reference corpora are typically the same size as the target corpus, or very much larger. Scott and Tribble (2006: 58) suggest that the reference corpus "should be an appropriate sample of the language which the text we are studying . . . is written in," and that "appropriate sample" "usually means a large one, preferably many thousands of words long and possibly much more." Precisely what counts as large enough is still a matter of debate. Scott's (2009) findings indicate that the content of the reference corpus is more crucial than the size, where only one register or text type is under investigation. Xiao and McEnery (2005: 70) compare two reference corpora, the 100-million-word BNC and the one-million-word Freiburg–LOB Corpus (see Hundt *et al.* 1998), and achieve almost identical keyword lists, thus concluding that "the size of reference corpus is not very important in making a keyword list."

1.5.2 Minimum frequency

The minimum frequency cut-off parameter can be used to exclude words that will be identified as unusual simply because they happen not to have occurred, or have occurred very infrequently, in either the dataset or the reference corpus. Proper nouns, for example, are often amongst these one-off occurrences. This is not to say that such phenomena – which are referred to as "hapax legomena" – are uninteresting (see, for example, Hoover 1999: chapter 4). The problem is that in a list of keyword results, mixing frequent items with very infrequent items often means mixing generalized phenomena with phenomena that are extremely localized, making an account of the keyword list problematic (see the following subsection for a statistical technique designed to reduce this problem). It is advisable to experiment with different minimum frequency cut-off points to minimize this problem, whilst ensuring that sufficient results are generated (if the dataset is small). A further point of good practice is to provide the raw frequencies of each keyword, in addition to its keyness value (log-likelihood or chi-square) when keyword lists are given.

1.5.3 Statistical significance

The significance test calculates the significance of the difference in frequency between the word in the target data and in the reference corpus. Some programs offer a choice of statistical tests (e.g. WordSmith Tools offers both chi-square and log-likelihood), while others do not (Wmatrix, for example, offers only log-likelihood). Culpeper (2009b: 36) conducts keyword analyses using both tests and discovers that the "same results were revealed with only minor and occasional differences in the ranking of keywords, differences which had no effect on the overall picture revealed by the

keywords." According to Scott (2013: 199), log-likelihood "gives a better estimate of keyness, especially when contrasting long texts or a whole genre against your reference corpus," although he also suggests that "where the notion of risk is less important than that of selectivity, you may often wish to set a comparatively low *p* value threshold such as 0.000001 (one in 1 million) . . . so as to obtain fewer keywords" (2013: 195–196). The *p* value is a number between 0 and 1 used in log-likelihood and chi-square tests. It indicates the probability of a key result occurring by chance; see e.g. Scott (2013: 203–204). Rayson (2003), evaluating various statistical tests for data involving low frequencies and different corpus sizes, favors the log-likelihood test "in general" and, moreover, a 0.01% significance level "if a statistically significant result is required for a particular item" (2003: 155). Further information on significance levels and cut-off thresholds is given on the website http://ucrel.lancs.ac.uk/llwizard.html, which also offers a log-likelihood calculator.

Alternative or additional statistical manipulations have been proposed. In brief, these are:

Bayes Factors. Wilson (2013) argues that the interpretation of the significance test in keyword analyses is often erroneous, notably because of the role of the null hypothesis. The *p*-value is a conditional probability: it is based on a range of statistics conditional upon the null hypothesis being true (i.e. there is no actual difference in the frequency of this word between the larger populations from which the samples are drawn). Wilson's solution is to use Bayesian statistics which focus on the weight of evidence against the null hypothesis (i.e. against there being no true difference between the populations from which the samples were drawn). One of the benefits of this technique is that it more accurately pinpoints the items that are highly key in the list.

The Cochran rule. Expected frequencies, as opposed to observed frequencies (i.e. the number of cases that actually occur), are calculated in the background by keyword programs. They are important because they influence statistical reliability; expected frequencies below 5 have been shown to render chi-square statistics unreliable (Oakes 2009: 165–166; Rayson *et al.* 2004b: 928). The Cochran rule (Cochran 1954) has been proposed as a method for eliminating items with expected frequencies with low statistical reliability, and has been extended by Rayson *et al.* (2004b) on the basis of numerous experiments with different corpus sizes, word frequencies, and statistical cut-off levels (see also Walker 2012). However, the advantages of the Cochran rule may vary according to the kind of analysis applied to keywords, and, as with Bayes Factors, at the present time it needs to be applied manually.

Bonferroni's correction. It is also possible to apply additional statistical methods to adjust the *p* value in order to control for the fact that a keyword list is a set of multiple statistical comparisons, the argument being that the

probability of a result occurring by chance increases with the number of comparisons performed. See for example Gries (2003a: 82–87), with regard to the Bonferroni correction, and for further discussion Gelman *et al.* (2012).

1.5.4 Dispersion

Uneven dispersion makes the interpretation of keywords difficult (Leech *et al.* 2001; Gries 2008). We have already discussed the notion of dispersion in Section 1.3, where we showed how a keyword plot gives a visual representation of dispersion. Eyeballing such a representation is not a precise method. Some scholars have proposed that the dataset and its reference corpus should be divided into parts, and then two-sample tests such as the *t*-test (Paquot and Bestgen 2009) or the Wilcoxon-Mann-Whitney test (Kilgarriff 1996) should be deployed.

We would concur with Baker (2004) that the best general approach to settings for keywords is to determine them by testing various possibilities and, in most cases, choosing a combination that results in: (1) a sufficient number of results to meet one's research goals, (2) a not overwhelming number of words to analyze, (3) an adequate dispersion of at least some keyword instances, and (4) any one-off or extremely rare word types being minimized. It may be possible for future research to produce more precise guidelines, though settings cannot be reduced to a simple mathematical formula for the reason that different research purposes and contexts have different requirements. Minimally, all studies should declare the settings that are deployed.

1.6 The state of the art

Keywords and other key items have become established in corpus linguistics as useful methods for identifying the lexical characteristics of texts. It is recognized that the notion of exactly what qualifies as key in any study is influenced by the settings and parameters of the program used, and by the comparator texts (in the reference corpus). Keyness in corpus linguistics is but the first statistical step in the analysis of texts. As Scott emphasises, keyness is context-dependent, and keywords "are pointers, that is all" (2010: 56).

Relatively new software tools, notably Paul Rayson's web-based WMatrix (Rayson 2003, 2008, 2009), enable users to annotate their datasets for both grammatical and semantic categories relatively easily and rapidly, and then to identify which categories are key. Studies include: Jones *et al.* (2004) on key POS categories in a spoken corpus of English for Academic Purposes, Afida (2007) on semantic domains in business English, and Archer *et al.* (2009) on semantic domains in Shakespeare's plays (see http://ucrel.lancs.ac.uk/wmatrix for further references on Wmatrix). We examine the added value of POS and semantic domain analyses to word-level keyness in Section 2.

The contributors to Bondi and Scott (2010) explore some current issues surrounding the theoretical nature of keyness, provide new perspectives on different aspects of keyness, and extend the existing boundaries of keyness methods in examining specific text-types. Stubbs, for example, discusses the relationship between the concept of "keywords" in corpus linguistics (as a textual feature) and ideas of sociocultural "keywords" (see Section 1.1), arguing that "[s]ocial institutions and text-types imply each other: they are different way of thinking about the same thing" (2010: 40). He points out that the findings from corpus studies of keyness address research questions in particular discourse contexts, but do not yet extend to bigger "cognitive and social questions" (ibid.).[1] Groom focuses on closed-class keywords in a corpus of academic history journal articles, arguing that they "may actually be preferable over their open-class counterparts as objects of corpus-driven discourse analysis because they offer much greater coverage of the phraseological data in a specialized corpus in both quantitative and qualitative terms" (2010: 73). Warren (2010) airs the matter of aboutness, using "concgrams" (Greaves 2009; see also Scott 2013: 332) and "aboutgrams" to investigate two different engineering texts, a process he derives from Sinclair (2006). Gerbig (2010) uses keywords as a starting point, and extends her analysis of a diachronic corpus of travel writing to key keywords and associates, extended lexical units, key phrases, and phrase frames.

Other recent developments include McIntyre and Archer's (2010) use of key semantic domains to investigate mind style in a play, and the partial automation of metaphor analysis, also using key semantic domains (Archer *et al.* 2009; Koller *et al.* 2008; Deignan and Semino 2010). Finally, Baker (2011) puts forward the concept of a "lockword," a potential means of addressing similarities between texts, bearing in mind that keyness is oriented to differences (Baker 2004: 349). Baker argues that lockwords are words with the most similar high frequency, statistically, across several corpora, and that these can be considered the opposite of keywords (2011: 73; see Taylor 2013 for further exploration of lockwords and the "locking" concept).

1.7 The value of going beyond keywords: key parts of speech and semantic categories

Rayson (2008: 543) puts forward two arguments for conducting key POS and semantic analyses *in addition* to keyword analyses. He argues, firstly, that they present fewer categories for the researcher to grapple with. Whilst it may well be the case that fewer keywords can be produced by tweaking program settings, it is also the case that by conducting key POS

[1] A recent linguistic pragmatics study by Bigi and Greco Morasso (2012) investigates keywords which trigger or activate cognitive frames in argumentation, using a psychologically salient concept of a "keyword," i.e. a sociocultural concept not a statistical concept.

and semantic analyses one can get clues to patterns that exist in a large set of keywords. Second, Rayson argues that POS and semantic categories can group lower-frequency words which might not appear as keywords individually and could thus be overlooked. In the case study in Section 2, we assess what added value the researcher can gain from the application of keyness methods to POS and semantic domains, in addition to single words, by presenting some data and analysis from Culpeper's (2009b) research.

2 Case study: evaluating key part of speech and semantic domain analysis

2.1 Introduction

Culpeper (2009b) addresses the following research question: by extending keyness analysis to grammatical or semantic tags, precisely what do we gain beyond what we could have learnt from a keyword analysis alone? Importantly, he sought to quantify what additional findings would be reached through the investment of time and effort in the further analytical steps of investigating key POS and key semantic domains. The results illuminate how and where these newer keyness techniques can most usefully be deployed. We present these here in order to assist researchers in making informed choices about the kinds of key linguistic structures that will most usefully serve their aims. The continuing development of keyness tools means that scholars have ever-widening choices to make about which techniques and programs to deploy from those that are now available, yet time and other resources are inevitably finite. Limited space means that we present a selection of results and analysis only, in Section 2.3, following an outline of the methodology used (in Section 2.2).

2.2 Methodology

Culpeper's (2009b) study built on the keyword analysis of Shakespeare's *Romeo and Juliet* reported in Culpeper (2002), using the same data, sourced from the 1914 Oxford edition of the plays edited by W. J. Craig. The data comprise the speech of the six characters who speak the most (their total speech varying from 1,293 to 5,031 words) plus reference corpora devised for each character containing the speech of all the others. Keywords were derived using *WordSmith Tools*. The statistical criteria used are a log-likelihood value of 6.63 or higher, which is equivalent to $p < 0.01$, and a raw frequency value of five or more. These same criteria are maintained for all the analyses in Culpeper (2009b). To illustrate, the positive (over-used) keywords derived for Romeo were (in rank-order of strength of keyness): *beauty* (10), *love* (46), *blessed* (5), *eyes* (14), *more* (26), *mine* (14), *dear* (13), *rich* (7), *me* (73), *yonder* (5), *farewell* (11), *sick* (6), *lips* (9), *stars* (5), *fair* (15), *hand* (11), *thine* (7), *banished* (9), *goose* (5), and *that* (84).

Culpeper (2009b) focused on three characters: Romeo, Mercutio, and the Nurse, whose keywords he determined to be characterized by different functions (see Halliday e.g. 1994): ideational (Romeo), textual (Mercutio) and interpersonal (the Nurse). Texts containing the speech of Romeo, Mercutio, and the Nurse were then uploaded into WMatrix, which automatically ran them through a number of programs. The Constituent Likelihood Automatic Word-tagging System (CLAWS) applies POS tags (see further Garside 1987; Leech *et al.* 1994). The UCREL (University Centre for Computer Corpus Research on Language) Semantic Analysis System (USAS) tool applies semantic tags, and is an annotation program designed for automatic dictionary-based content analysis (see Rayson *et al.* 2004a).[2] Needless to say, neither POS nor semantic tagging achieves a perfect result (semantic tagging is claimed to achieve an accuracy rate of 91 percent with present-day English by Rayson *et al.* 2004a). Each and every resulting key item was checked manually, in order to assess whether or not that item accounted for a textual pattern or style that had also been accounted for in the other analyses (and also, in the case of the grammatical and semantic analyses, whether or not that item was simply a tagging error).

2.3 Results of key part of speech analysis

Below, we briefly introduce some of the POS and semantic category results and discussion for Romeo (based on Culpeper 2009b). Table 5.1 displays

Table 5.1 *Romeo's parts-of-speech rank-ordered for positive keyness (i.e. relative overuse) (keywords are in bold text)*

Grammatical category (and tag code and frequency)	*Items within the category (and their raw frequencies) up to a maximum of ten types if they are available (excluding clear tagging errors in square brackets)*
Nominal possessive personal pronoun (e.g. *mine, yours*) (PPGE) (17)	**mine** (8), hers (4), **thine** (3), [his (1)], yours (1)
Comparative after-determiner (e.g. *more, less, fewer*) (DAR) (16)	**more** (15), less (1)
1st person sing. objective personal pronoun (i.e. *me*) (PPI01) (73)	**me** (73)
General adjective (JJ) (328)	**fair** (14), good (10), **dear** (10), sweet (8), **rich** (7), dead (6), holy (5), true (5), heavy (5), **blessed** (4)
1st person sing. subjective personal pronoun (i.e. *I*) (PPIS1) (144)	I (144)
Than (as conjunction) (CSN) (16)	than (16)

[2] Both CLAWS and USAS were developed at Lancaster University. More information about the CLAWS tagger can be found at: www.comp.lancs.ac.uk/ucrel/claws, and further details about the USAS tagger at http://ucrel.lancs.ac.uk/usas/.

the grammatical categories that are key in Romeo's speech. Insecure items, mostly because of tagging errors, are in square brackets.

With the exception of the category General adjective (JJ), all the key grammatical categories in Table 5.1 are dominated by a single item. This is because such categories have high frequencies of tokens but a relatively limited range of types. The category General adjective (JJ) stands in contrast, because it is a more contentful lexical category including a broader and more even group of items. The more grammatical categories tend to contain items which have already occurred in the keyword results, since those categories are mainly populated by high-frequency items that are also likely to appear in a keyword analysis. Four of the most frequent words for each of the six categories (*mine*, *more*, *me*, *fair*) also occur as keywords. Note that among them is the category General adjective (JJ), although the status of this whole category is questionable because most of the instances of *good* and *dear* occur in vocative expressions. In contrast, in the more grammatical categories, the first-person singular subjective personal pronoun (*I*; PPIS1) and *than* (as conjunction; CSN), do *not* include members which are keywords at the significance level used in this paper (i.e. $p < 0.01$).[3] Therefore, we cannot say that the prevalence of a limited number of high-frequency items in grammatical categories means that they will also occur in a keyword list, although this tends to be the case.

Table 5.2 displays the semantic categories that are key in Romeo's speech. In Table 5.2, very few lexical items in the semantic categories also occur as keywords. The top two categories, Relationship: Intimate/sexual (S3.2) and Liking (E2+), are, of course, closely linked semantically. Considered overall, the appearance of these categories as most key is very well motivated; as we might expect, the findings confirm Romeo's role of one of the lovers in the play. The contents of the third-ranked category, Colour and colour patterns (O4.3), are rather less predictable. In some cases Romeo describes literal light, e.g.: "But, soft! what *light* through yonder window breaks?" (II.ii). However, the terms are more often used metaphorically, e.g.: "More *light* and *light*; more *dark* and *dark* our woes" (III.v). Such metaphors are quite conventional: *light*/*dark* to mean happiness/unhappiness, *greenness* to mean envy, and *redness*/*whiteness* to mean life/death. The semantic tagger does not (currently) make a distinction between literal and metaphorical meanings. Despite this, metaphorical patterns surface in some of the semantic categories which occur as key.

To reach a more definite general conclusion about what is to be gained from extending a keyness analysis to POS categories or semantic

[3] In the keyword analysis, the word forms *I* scores and *than* achieve log-likelihood of 5.52 and 6.18, respectively, both of which are lower than the critical value of 6.63 for $p < 0.01$. These single items may occur as key in grammatical categories but not as key*words* because the more items that populate a category, the harder it is for differences in frequency between items in that category to surface. Differences relating to a specific item in that category could be averaged out by other items.

Table 5.2 *Romeo's semantic categories rank-ordered for positive keyness (i.e. relative overuse) (keywords are in bold text)*

Semantic category (and tag code and frequency)	*Items within the category (and their raw frequencies) up to a maximum of ten types if they are available (excluding clear tagging errors in square brackets)*
Relationship: Intimate/sexual (S3.2) (48)	**love** (34), kiss (5), lovers (3), kisses (2), paramour (1), wantons (1), chastity (1), in love (1)
Liking (E2+) (38)	**love** (15), **dear** (13), loving (3), precious (2), like (1), doting (1), amorous (1), [revels (1)], loves (1)
Colour and colour patterns (O4.3) (33)	light (6), bright (4), pale (3), dark (3), green (2), stained (2), black (2), golden (1), white (1), crimson (1)
Education in general (P1) (9)	teach (3), [course (2)], philosophy (2), school (1), schoolboys (1)
Business: Selling (I2.2) (19)	sell (4), [bid (4)], shop (2), hire (2), buy (1), sold (1), [stands (1)], [bade (1)], [stand (1)], [store (1)], merchandise (1)
Thought, belief (X2.1) (26)	think (7), feel (3), devise (2), believe (2), [take thence (1)], thinking (1), thought (1), engrossing (1), dreamt (1), [found (1)], in thine eyes (1), in mind (1)
Affect: Cause/Connected (A2.2) (20)	[hence (7)], reason (2), [spurs (2)], depend (1), for fear of (1), provoke (1), excuse (1), effect (1), consequence (1), to do with (1), appertaining (1), prompt (1)
Avarice (S1.2.2+) (7)	envious (3), [mean (1)], tempt (1), jealous (1), sparing (1)
The universe (W1) (21)	world (8), [word (6)], **stars** (5), moon (2)
Money: Affluence (I1.1+) (7)	**rich** (7)

categories, Culpeper (2009b: 53–54) quantifies the benefits of each type of analysis for the three characters. Although the raw frequencies are low, in percentage terms 75 percent of the POS categories are mainly populated by one or two words which also occur as keywords, and 66.6 percent of the semantic categories mainly constitute one or two words which also occur as keywords. Therefore, although the tests clearly need to be replicated on bigger datasets, the findings indicate that a key*word* analysis led to most of the conclusions. At face value, the difference between 75 percent and 66.6 percent appears to be of little or no consequence. It is notable, though, that the POS and semantic domain results for Romeo overlapped rather less with the keyword analysis: 66 percent for the POS categories and 40 percent for the semantic categories. When keyword lists are mainly populated by those which are ideational, which reflect the 'aboutness' of the text, the POS, and particularly the semantic keyness analyses can make a more valuable contribution, taking the analysis to a level beyond that which emerges from the keywords. The likely explanation for this is that

more grammatical items, and also discourse markers, tend to occur as a relatively limited range of types, each with frequent tokens. This means that, if categories constituting these kinds of items are identified as key in the POS or semantic analyses (as they are for the Nurse and Mercutio), then such items are also very likely to occur in the keyword analysis.

The outcomes of Culpeper (2009b) lend support to Rayson's (2008: 543) arguments for conducting key POS and semantic domain analyses *in addition* to keyword analyses, given at the start of this section. In particular, Rayson's claim that POS and semantic categories can group lower frequency words, which might not appear as keywords individually and could thus be overlooked, is confirmed in the analyses included in this chapter. Examples include general adjectives and Romeo's (metaphorical) colour terms. These kinds of results are not easy to predict, but closer analysis indicates that they are well motivated. It should be noted, though, that the added value obtained from these analyses is more specific than indicated by Rayson, as it applies only to more lexical, more ideational categories.

3 Conclusion

In addition to charting the history and development of keyword research (in Section 1), our discussions have been designed to emphasize that key lexical items should be used as a *guide* for what to analyze qualitatively, and not considered the end product in themselves. We have emphasized (particularly in Section 1.5) that the usefulness of key items, and the quality of analyses and conclusions based upon them, relies on careful and explicit manipulation of the keyword tools settings as well as interpretation. Finally, our case study in Section 2 aimed to demonstrate that, given the availability of new and varied techniques for investigating keyness in different kinds of linguistic structures, it is necessary to consider and test out which one(s) will most usefully target the language features the researcher wishes to uncover.

The future of work involving keyness looks bright. We can expect developments in three directions. First, there is scope for methodological improvements. As discussed in Section 1.5, statistical refinements have been suggested, some of which have yet to be integrated into mainstream programs (e.g. log ratio as a means of taking effect size into consideration in the ranking of keyword results is being incorporated into a number of programs). Second, the extension of keyness analysis to POS and semantic categories need not stop there. In principle, any computer-readable form or user-supplied tag could be interrogated. For example, styles of punctuation could be thus investigated. Third, we are only just beginning to see the deployment of keyness analyses in academia (currently, many studies have focused on literary texts). The full potential of keyness analyses across the humanities and social sciences has yet to be realized.