LANGUAGE IN THE NEWS: SOME REFLECTIONS ON KEYWORD ANALYSIS USING *WORDSMITH TOOLS* AND THE BNC

Sally Johnson & Astrid Ensslin

Abstract

It is not uncommon to hear linguists lamenting the *mis*-representation of language whenever linguistic issues are taken up by the media. Ironically, however, we have relatively little systematic understanding of the ways in which language is actually dealt with in, and by, those media. This paper focuses on methodological issues that arose in the context of a project that aimed to explore the ways in which themes relating to language and linguistics are represented in a corpus of articles gathered from two British newspapers, *The Times* and *The Guardian*. Using the software programme *WordSmith Tools* (Scott, 2004) to identify those 'key' keywords that were most likely to occur in conjunction with the 'node terms' <language>, <languages>, <linguistic> and <linguistics>, this corpus-based methodology revealed a number of interesting ways in which language-related issues are debated in this particular sector of the print media. At the same time, as will be discussed in this paper, our study raised some important methodological concerns in relation to the use of *WordSmith Tools*, the British National Corpus, and the construction of keyword lists.

1. Introduction

It is widely acknowledged, and frequently lamented, within the discipline of linguistics that the views of linguists as language 'experts' are not always accorded a high degree of respect outside of academe (e.g. Aitchison, 1997; Bauer & Trudgill, 1998; Laforest, 1999; Rickford, 1999; Johnson, 2001). It has often been suggested therefore that if the academic study of language is to be of greater relevance to public practice, it is imperative that we try to understand, as linguists, the epistemological conflicts that can arise between 'expert' and 'lay' conceptualisations of language-related phenomena (e.g. Antos, 1996; Bygate, 2004; Cameron, 1995, 2000; Niedzielski & Preston, 2000). Over the past few years, a number of contributions to the *Journal of Sociolinguistics* have highlighted the *media* as a key site for the generation of potential misunderstandings and misrepresentations where questions of language and linguistics are concerned (Wolfram, 1998; Cyr, 1999; Heller, 1999a, 1999b; Laforest, 1999; Rickford, 1999; Aitchison, 2001; Johnson, 2001; Milroy, 2001; Garrett, 2001). Such debates over the role of the media in both representing and constructing language phenomena, together with the often tense relationship between linguists and media producers, can in turn be situated in the context of a burgeoning literature on language ideologies (Blommaert, 1999; Gal & Woolard, 2001; Kroskrity, 2000; Lippi-Green, 1997; Schieffelin et al., 1998) together with the study of metalanguage as constituted by, and constitutive of, linguistic reality (Jaworski et al., 2004).

There is already a sizeable body of work that has begun to explore the representation and/or construction of language-related issues within a variety of media texts, both in terms of metalinguistic commentary (where the question of language is overtly thematised) and metapragmatic dimensions (where language is itself used in certain ways with particular discursive/communicative effects) (see Jaworski et al., 2004; Ensslin & Johnson, forthcoming). What broadly unites such work in theoretical terms is a concern to explore how language is itself thematised and/or deployed in media texts such that we might further our understanding of the ways in which particular conceptualisations of language varieties and their communities of users or 'publics' (Gal & Woolard, 2001) are normalised, thereby helping to construct, maintain, and sometimes even resist hegemonic 'regimes of language' in the context of dominant social power relations (Kroskrity, 2000). In this sense, media discourses form part of a much broader set of discursive processes within what Jan Blommaert (1999) refers to as 'language ideological debates' – debates that typically cluster around such themes as the status of standard, non-standard, and minority language varieties, together with the construction of user communities in relation to local, national, global and/or hybridised forms of identity.

In the project to be described in this paper, our aim was to contribute to the growing discussion of the representation/construction of language in media texts by adding a quantitative, corpus-linguistic dimension to the question of how language-related themes are conceptualised in the British broadsheet landscape represented, in this instance, by *The Times* and *The Guardian*. More specifically, our sociolinguistic research questions revolved around the construction of gendered language use (Johnson & Ensslin, forthcoming) and 'Englishness' (Ensslin & Johnson, forthcoming). Print media texts provide both a particularly useful and widely accessible source of data for observing and analysing such language ideological processes (see DiGiacomo, 1999: 105). With this in mind, our project was intended to illustrate how work in this particular field might in fact be enhanced by the use of large-scale electronic corpora and the innovative software that can nowadays be employed in their analysis. In this way, we wanted to introduce a more explicitly quantitative dimension to the study of a larger body of data than is typically found in this particular area of sociolinguistics and language ideology studies, thereby illustrating how such methods might be fruitfully combined with more traditional qualitative approaches. A main concern of this paper is therefore a description and evaluation of the methodology we employed in relation to the construction and analysis of our corpus. In so doing, we adopt a critical view towards the otherwise widely commended (e.g. Hunston, 2002; Johnson et al., 2003; Kemppanen, 2004; Baker et al., 2006) keyword function provided by *WordSmith Tools 4.0* (Scott, 2004). As a matter of fact, Baker (2004) is one of the very few corpus researchers so far to have offered any kind of critique with regard to *WordSmith* methodology. He warns of a lexical-only approach which runs a risk of blurring degrees of salience and, consequently, misleading the actual research focus. Instead, he suggests a triangulation methodology which combines quantitative and qualitative investigations, e.g. comparisons between different sets of data, analyses of key clusters, grammatically and semantically annotated data, collocations and concordances.

Following on from Baker's approach, we will present and discuss our own experiences with the construction of lists of 'key' keywords using *WordSmith Tools* in relation to a comparator corpus[1] based on the list of keywords from the British National Corpus (BNC). Most importantly, we will be exploring in greater depth some of the methodological problems we encountered in the context of this process at the nexus between quantitative and qualitative data analysis and the theoretical consequences for the thematic focus of our study.

2. Methodology

Our study focuses on a corpus of texts gathered from two British broadsheet newspapers, *The Times* and *The Guardian*, published during the period 1 July 2004 to

[1] In this paper the terms 'reference corpus' and 'comparator corpus' are used interchangeably.

30 June 2005, the analysis of which forms the first stage of a larger-scale project exploring representations of language in the print media.[2] Although we recognise that it is difficult to state definitively the political orientation of any given newspaper, and because the actual breadth of political coverage within the British press is continually open to debate, we decided to focus on these two particular newspapers, which are commonly classified as 'of the right' (*The Times*) and 'of the left' (*The Guardian*), and hence representative of the British broadsheet landscape. Archives for both newspapers were readily accessible and available free of charge via the online database *Newsbank UK*.

After selecting the newspapers and specifying the time-frame, the next step was to conduct an electronic search for the following four 'node terms': <language>, <languages>, <linguistic> and <linguistics>. Each article in which one or more of the terms in question appeared was then saved as an electronic text file. ('Repeat' texts were subsequently removed where a node term occurred more than once within a given sub-corpus but one copy retained within each of the four sub-corpora, where applicable). The overall search resulted in the compilation of four sub-corpora for each of the node terms, which we collectively refer to as LANGCORP. This 5 million-word corpus consists of approximately 7,000 texts containing just under 10,000 tokens of the four terms in question, as summarised in Table 1. N.B., it was not our aim to juxtapose the two newspapers and their specific uses of the four search terms. This would have been interesting for a contrastive media-ideological study, which we did not pursue. Instead, we aimed to explore the ways in which themes relating to language and linguistics are represented longitudinally across two representative British broadsheets.

Table 1 – Occurrences of four terms across the two newspapers

Node term	*The Times*	*The Guardian*	Total in LANGCORP
Total number of tokens	2,317,481	3,161,506	5,478,987
<language>	3,501	4,381	7,882
<languages>	668	850	1,518
<linguistic>	199	177	376
<linguistics>	34	55	89
Total number of search terms	4,402	5,463	9,865

Once the corpus was constructed, the next step was the identification and analysis of those 'keywords' that were most typically associated with the four terms in question. For this, we adopted the approach used by Johnson et al. (2003) in their study of keywords relating to 'political correctness' as well as Norman Fairclough's (2000) analysis of the language of New Labour.

Our understanding of the notion of keywords is derived from the work of Mike Scott (1997, 1998, 2000, 2001a, 2001b, 2002; cf. Stubbs, 1996), which has been widely applied and further developed, for instance by Ku & Yang (1999), Kemppanen (2004), Tribble (2000) and Scott & Tribble (2006). Unlike Raymond Williams, who

[2] We are grateful to the British Academy for providing the funding, which enabled us to carry out the work presented here in the summer of 2005 (ref. SG – 39228).

in his now classic *Keywords* of 1976 focuses on individual words and their cultural and ideological significance in general (cf. Firth, 1957), for Scott, words are ‘key’ in relation to the specific stretch of written or spoken discourse of which they are a part. Keywords are therefore identified by comparing word frequencies within a text or corpus of texts with word frequencies of another (usually much larger) reference corpus. According to Scott, keywords can reveal the ‘aboutness’ or content of a text, whereby the notion of ‘aboutness’ is attributed to Phillips (1989). One might also add here that Scott's understanding of keywords overlaps to a degree with what scholars in stylistics have referred to as ‘style-markers’, which come about when there is a significant differential between the densities of linguistic features in a text and the densities of corresponding linguistic features in a *contextually-related* norm (see Enkvist, 1964; 1973).

Scott’s computer program *WordSmith Tools 4.0* (2004) allowed us to conduct the statistical analysis that is required in order to generate a list of ‘key’ keywords.[3] According to Scott, ‘a word will get into the listing here if it is unusually frequent (or unusually infrequent) in comparison with what one would expect on the basis of the larger wordlist’ (Scott, 2004: Help Menu).[4] At this juncture, a key consideration was the nature of the reference corpus to be employed. Here we followed the procedure frequently advocated and adopted by other analysts (e.g. Tribble, 2000; Scott, 2000; 2001b; 2002; Johnson et al., 2003; cf. Hunston, 2002) and used the British National Corpus (BNC) as the reference corpus. Specifically, we used a word-list based on the entire BNC set of written texts (90.7 million words) taken from the late 1980s and early 1990s, a list that was constructed by Scott and is readily obtainable via his web-page (http://www.lexically.net/wordsmith/; accessed 2006/01/11).

In an attempt to gain a sense of the wider discourses in which the four node terms <language>, <languages>, <linguistics> and <linguistic> were embedded, we proceeded to generate a separate keyword list for each. On the basis of procedures identified elsewhere (see references and footnotes), we then filtered out the following items:

a) Word forms that reflect newspaper discourse in general such as ‘is’, ‘has’, ‘who’ and ‘says’; [5]

b) Word forms that belong to what McLachlan and Reid (1994: 104) refer to as part of the ‘circumtext’ of a given article, such as ‘author’ (name of the author), ‘paper’, ‘title’, ‘Guardian’ and ‘Times’ (newspaper title), ‘London’ and ‘England’ (place of publication), ‘date’ (of publication), ‘section’ (rubric), ‘page’, ‘copyright’, as well as

[3] To derive the ‘key-ness’ of an item, first the program computes (a) its frequency in the research text or corpus in question, (b) the number of running words in the research text/corpus, (c) its frequency in the reference (comparator) corpus, and (d) the number of running words in the reference corpus. Then it cross-tabulates these, and applies either of two statistical tests: the chi-square test of significance with Yates correction for a 2 X 2 table or Ted Dunning’s Log Likelihood test (see Scott 2004: Help Menu; Tribble, 2000: 79-80).

[4] The keywords we discuss in this paper are all what are known as ‘positive’ keywords, that is to say, they are ‘key’ because they are unusually frequent. This is partly because of space limitations in this paper, but also because (a) ‘negative’ keywords (which are ‘key’ because of unusual infrequency) tend to be very few in large data sets such as ours, and (b) the settings in WordSmith Tools are in any case biased towards positive keywords (see Tribble, 2000: 80-84).

[5] See Johnson et al., 2003: 35. In addition, we noted that two of the items, ‘who’ and ‘says’, were found by Biber et al. (1999: 375, 610) to be outstandingly frequent in newspaper language generally.

terms deriving from set phrases such as ‘newspapers’, ‘limited’, ‘all’, ‘rights’ and ‘reserved’;

c) Alternative grammatical forms of the same lemma, which occurred more than once amongst the top 30 keywords (usually in singular or plural form). Here we only retained the term that occurred first in the keyword list whereby such cases are marked with an asterisk (e.g. ‘university’, ‘schools’ and ‘teacher’);[6]

d) Proper names of central public figures, which Scott excludes from ‘true’ keywords, as they are specifically characteristic of media discourse during the time that is reflected by the research corpus and therefore unlikely to occur to a similar degree in closed, asynchronous reference corpora such as the BNC (see Scott, 2000: 115);

e) Terms relating to recent technological innovations such as the World Wide Web which appeared considerably more frequently in LANGCORP than in the BNC due to the difference in construction dates. The BNC was completed in 1993, which predated the upsurge in new media phenomena and terminology, such as ‘WWW’, ‘Google’, and ‘.com’, from the mid-1990s onwards;[7]

f) Word forms which only occurred (in disproportionately high numbers) in one single article or type of article that was considered statistically and conceptually insignificant (e.g. lists of election results after the general elections in Britain in May 2005 and the United States in November 2004; personal columns). This practice is justifiable or even sensible on the grounds of the homogeneity postulate as expressed by Sinclair (2004). Sinclair refers to texts that differ radically from the others in a corpus and, hence, threaten to distort the data for the sake of overly objective coverage, as ‘rogue’ texts. To guarantee accountability, he adocates the documentation of such exclusion procedures.

The results of the keyword analysis for each of the four terms are contained in Tables 2-5.

Table 2: Top thirty ‘key’ keywords for <language>

1	English	16	book
2	film	17	novel
3	anti	18	American
4	school	19	university
5	self	20	culture
6	non	21	story
7	music	22	students
8	his	23	poetry
9	BBC	24	immigration
10	pre	25	director
11	world	26	writer
12	theatre	27	born

[6] Clearly, if precisely those keywords were to form the object of close analysis, instances of both word forms would have to be taken into account.

[7] This practice was reconfirmed by Mike Scott and Paul Rayson during a corpus-methodological debate at the 2006 AHRC IT Methods Network ‘Historical Text Mining Workshop’, Lancaster University (20th-21st July 2006).

13	services	28	life
14	year	29	media
15	French	30	war

Table 3: Top thirty ‘key’ keywords for <languages>

1	university*	16	pupils
2	schools*	17	GCSE
3	students	18	writer
4	English	19	teacher*
5	education	20	fiction
6	French	21	maths
7	non	22	life
8	bursaries	23	self
9	born	24	translated
10	novel	25	Spanish
11	year	26	career
12	anti	27	culture
13	teaching	28	world
14	books*	29	Latin
15	college	30	modern

Table 4: Top 30 ‘key’ keywords for <linguistic>

1	English	16	bling
2	non	17	Spanish
3	Piraha	18	monkeys
4	anti	19	Iraq
5	self	20	sandwich
6	Ukraine	21	Arabic
7	university	22	nursery
8	French	23	Mullan
9	fiction	24	slang
10	Derrida	25	Palestinian
11	word*	26	Gaelic
12	novel*	27	polyamory
13	book	28	QCA
14	Truss’s	29	media
15	poetry	30	narrators

Table 5: Top 30 ‘key’ keywords for <linguistics>

1	professor	16	studies
2	university*	17	reading
3	Polari	18	lecturer
4	English	19	Cairo
5	PhD	20	non
6	biological	21	postgraduate
7	students	22	history
8	graduate*	23	academic
9	Sudan	24	Anglo
10	phonetics	25	engineers

11	school	26	biology
12	science	27	Egyptian
13	Arabic	28	translation
14	spelling	29	Tolkien
15	engineering	30	gay

A preliminary analysis of the four keyword lists shows how the lemmas *English*, *university* and *non* occur in all four lists. In more general terms, it is possible to identify the four dominant semantic fields listed in Table 6.

Table 6: Four semantic fields and related types

Semantic fields	Types
Language(s)	'English', 'French', 'translated', 'Latin', 'modern', 'Polari'[8], 'Arabic', 'translation', 'Piraha'[9], 'Gaelic'
Education	'school(s)', 'university/universities', 'students', 'education', 'bursaries', 'teaching', 'college', 'pupils', 'GCSE', 'teacher(s)', 'maths', 'PhD', 'graduate(s)', 'phonetics', 'science', 'spelling', 'studies', 'lecturer', 'postgraduate', 'academic', 'history', 'biology', 'nursery', 'QCA' ['Qualifications and Curriculum Authority']
Media culture	'film', 'music', 'BBC', 'theatre', 'book(s)', 'novel(s)', 'culture', 'story', 'poetry', 'writer', 'media', 'fiction', 'reading', 'Tolkien', 'narrators'
Identity	'self', 'immigration', 'born', 'life', 'culture', 'world', 'Anglo', 'Egyptian', 'gay', 'Palestinian', 'polyamory'

Also of interest is the repeated occurrence of the bound morphemes and/or lexical prefixes 'non', 'anti', 'self' and 'pre', which tend to be used for the purposes of signalling anteriority, on the one hand, or, more prominently, the negation or resistance of a particular identity, on the other. Here the three keywords 'non', 'anti' and 'self' occurred in the keyword lists for <language>, <languages> and <linguistic>, whereas 'pre' occurred only in the sub-corpus for <language>. This collocational frequency suggests a strong interrelationship between linguistic items representing language-related issues and those representing identity-related ones, a hypothesis which, for its conceptual validity, would need to be investigated by means of concordance analyses and close contextual readings.

Whilst such a preliminary analysis on the basis of quantitative procedures can give us some indication of the semantic patterning of the corpus, a more in-depth qualitative investigation is, of course, required if we are to be able to explore the ideological implications of the text corpus. On the basis of the four keyword lists, it is possible to develop hypotheses that can then be explored in greater depth using concordance analyses (Hunston, 2002: 55). That said, it goes without saying that the sheer size of the corpus (and indeed most corpora) will not permit the analysis of all thematic possibilities. Elsewhere we have therefore focussed on an exploration of the relevance of 'Englishness' in relation to language, based on an analysis of texts containing the 'key' keyword 'English' which occurred within the top four positions in all four keyword lists (Ensslin & Johnson, forthcoming). Moreover, we have also

[8] See Baker (2002).

[9] *Piraha* refers to an indigenous Brazilian group together with its language.

looked in depth at the representation of gendered styles of language usage based on texts containing the string <his language> and supplemented by a small number of occurrences of texts containing the string <her language>, an analysis that was initially prompted by the appearance of the 'key' keyword 'his' in position number eight on the list for <language> (Johnson & Ensslin, forthcoming). What we wish to do in the remainder of this paper, by contrast, is to explore some of the specifically methodological issues that we encountered in the process of conducting our analyses. These, in turn, raise concerns that are then directly of relevance to the more general theoretical aims of our study, namely to explore the ways in which themes relating to language and linguistics are represented in the two newspapers concerned.

3. Methodological concerns

Our main interest with respect to keyword analysis lay in the insight we sought to gain into unusually frequent lexical items in LANGCORP compared to our reference corpus, which was intended to represent 'default' language use. We expected the keyword lists to enable us to draw conclusions for ensuing qualitative (discourse) analyses, which would be based on concordance lists and collocational patterns. In what follows, we will flag up two specific methodological concerns which we found ourselves confronted with in the context of our quantitative examinations.

3.1 The use of the BNC as comparator corpus

The choice of reference corpus is a major question in any corpus linguistic project that seeks to identify (key) keywords in a particular research corpus. As previously mentioned, in our project, using the BNC as comparator corpus seemed to be the most feasible, if not the only available option. Firstly, use of the BNC in such contexts as these has been documented as common practice by a number of leading researchers and keyword analysts in the field. Scott (2001b: 126), for instance, claims that the BNC is adequate in that the effect of any discrepancies between two corpora will be minimal. In fact, Scott emphasises that 'in practice [he does] not find that changing the reference corpus makes much difference' (2000: 115). Secondly, building a reference corpus of approximately one-hundred million words (at least ten times the size of the research corpus) from scratch, which would have truly allowed us to compare 'like with like', would have presented resource implications well beyond the scope of our study (and presumably many others like it). However, a range of observations made during our study leads us to wonder just how useful and relevant the BNC really is as a comparator corpus for keyword analysis using latter-day corpora.

As a matter of fact, the choice of the BNC as reference corpus in any recent, synchronic corpus project (i.e. any project recording immediately contemporary language use that was started less than five years ago) is likely to confront researchers with a number of issues that will have a major impact on the research design. Firstly, and perhaps most prominently, the fact that the BNC, or indeed any other closed comparator corpus, was completed a relatively long time (over a decade) before the research corpus in question causes the problem of age disparity. In other words, the question is whether certain themes (and the word forms represented by them) that come up in a keyword list are genuinely more salient in a research corpus or whether they just do not (yet) occur in the reference corpus. Such 'themes' comprised, in our case, a large number of word forms from the field of computing, the World Wide Web and virtual reality, which, in 1993, when the BNC was completed, did not yet form part of media discourse to quite the same extent as they do today. The question is therefore whether any, or indeed only specific, types of references to internet-related concepts are to be dismissed as part of an article's circumtext in the same way as references to, say, 'copyright'. Alternatively, those themes may not be considered central to the thematic focus of our particular study at all insofar as we are concerned with language. Furthermore, internet-related concepts and word forms constitute part of media discourse in general nowadays in a similar way as the previously-mentioned word forms 'is', 'has', 'who' and 'says'.

That said, one cannot possibly exclude concepts relating to the internet, the WWW or any new electronic media in general from any study of language-related discourse such as ours, given that they are closely associated with, or perhaps even form *the* major site of metalinguistic commentary nowadays. We pursue this idea further in the next section, which adds to our discussion the problematic dimension of proper names in newspaper and media discourse.

3.2 Proper names

Closely related to the previous issue is the problem of proper names. Naturally, in any newspaper corpus, proper names of public figures, for instance from politics, the economy, sports, and the entertainment industry will appear as 'key' keywords in a number of texts if they have dominated media discourse over a certain period of time, provided an asynchronous comparator corpus, like the BNC, has been used. Scott (2000: 115) categorically excludes proper names of any kind (including authors' names) from lists of 'true' keywords on account of the age disparity problem. Nevertheless, politicians' names in particular are often closely associated with topical themes that are directly relevant to the research objectives (in our case, looking at representations of language- and linguistics-related issues, such as debates over multilingualism, held during the 2004 US elections). Furthermore, with the exception of household names such as 'Blair', 'Chirac' and 'Beckham', the majority of proper names only occurred in a statistically insignificant number of texts and hence cannot be considered 'key' keywords. In fact, Sinclair (2004) has argued that such articles should actually be excluded because they endanger the homogeneity and hence representativeness of the research corpus.

Similarly, our specific *theoretical* problem here, which goes back to the aims of the study, relates to the question of 'ideological brokers' as identified by Blommaert (1999) and DiGiacomo (1999). In a critical sociolinguistic study such as ours, the central conceptual research interest lies in highlighting the 'real social actors' who are engaged in debates over language and linguistics, and in analysing how their discourse is represented in the media. We are thus dealing with a double-layered analysis which is equally interested in prominent public figures' (direct) discourse regarding language-related issues, and the ways in which this discourse is represented by means of media-governed *meta*-discourse, which again casts light on immanent media ideologies.

In all, any keyword analyst or, for that matter, any comparative corpus linguist will have to undertake complex decision-making processes, which involve, on the one hand, the careful selection or indeed construction of an appropriate comparator corpus and, as a potential result, the partial or entire inclusion and/or exclusion of certain semantic fields, lexical items and word forms according to their relevance to the research objectives, on the other. Clearly, a certain degree of editorial work cannot be avoided when it comes to deriving a 'true' keyword list from the 'rough' keyword list produced by the software. However, in a large corpus this 'clearing' process can only be done effectively if it is done manually and all the texts are looked at individually, which, after all, is not really what digital corpus work ought to be concerned with.

4. Discussion and concluding remarks

As indicated above, the prevailing issue regarding the selection of an appropriate reference corpus is contingent upon the degree to which researchers need to be able to compare like with like, which, as a matter of course, varies from research context to research context. In this respect, the question of what 'like with like' refers to is not only limited to the decision between a pre-existing or self-constructed reference corpus. In our particular case, there were, for instance, considerations of genre (concerning the types of texts, rubrics or articles to be included in a self-constructed reference corpus), source (broadsheet newspapers only or a combination of broadsheets and tabloids), mode (written or spoken language), and time frame (exactly synchronous, partly or fully asynchronous, and, in case of the latter, the precise temporal difference, which, of course, would have to be accounted for).

The methodological problems described in the previous section seem to suggest that the only way forward is to build one's own comparator corpus and construct an individual keyword list. Most researchers would undoubtedly feel this to be impractical. We considered the possibility but dismissed it a) on the grounds of time involved and the limitations it would create in terms of the analysis and b) because the common procedure, well referenced in the literature, is to use the BNC. In hindsight, however, our chosen practice resulted in unforeseen editorial efforts, due to the unexpected amount of time that had to be invested in clearing the keyword lists, and in the above-described complexities arising from selective decision-making.

Evidently, the methodological problems described in this paper would have been solved by building our own exactly synchronous comparator corpus consisting of all of the texts within the two newspapers for the said period of time. Firstly, the software would have generated keyword lists that, although some editorial work regarding circumtextual elements would still have been necessary, would have generated a more reliable 'rough' list of unusually frequent items. From this list, considerably fewer proper names would have needed to be deleted, as, with the exception of authors' names, they would have been clearly identified as contemporary, language- and linguistics-related 'ideological brokers'. Furthermore, recent computing jargon would, in all likelihood, not have occurred in the first place, and if it had, those keywords would have had to be submitted to closer investigation along similar lines to our analysis of the function word 'his', which was listed amongst the top thirty keywords in the <language> subcorpus and therefore gave rise to questions regarding representations of gendered language use (Johnson & Ensslin, forthcoming).

Undoubtedly one other possible solution to all of the points mentioned above is to conduct extensive editing work on the keyword lists, making individual decisions on what should remain or be excluded in line with thematic and theoretical focus of the study itself. After all, a certain degree of editing has to be done at any rate, as many non-lexical, non-grammatical items, such as numerals, orthographic signs and irrelevant abbreviations occur at the top of any keyword list, no matter how synchronous the reference corpus. However, whilst we do not believe that any one, least of all Scott, would ever claim that the keyword function can produce entirely objective results, manipulating computer-generated lists in this way feels intuitively problematic. Having turned to a corpus methodology and the keyword function precisely in order to add an element of objectivity and bottom-up processing to our own particular study, excessive individual editing of the lists does not appear to accord with those aims. In this sense, it is difficult to take corpus linguistic insights 'back' to the community of sociolinguists and linguistic anthropologists working in this area in order to be able to claim a methodological or theoretical breakthrough. It is also worth highlighting that critical discourse analysis, which constitutes a major qualitative method of investigation in this field of linguistics, is, *per definitionem*, a subjective process and therefore consistently susceptible to accusations of over- or under-interpreting 'objectively' derived data.

In view of these conflicting methodological concerns, we return by way of conclusion to Baker's (2004) advocacy of carefully triangulated quantitative and qualitative analytical methodology, which has to be seen as the foundation for non-structuralist corpus research. Clearly, when using a statistically grounded keyword methodology, looking at lexical evidence alone will never suffice for functional investigations. Therefore one has to exploit the software to include larger individual and textual patterns by means of collocational and concordance analyses and then combine the statistical findings with a range of 'inclusive and subjective' (Baker, 2004: 352) interpretations. In sum, in order that we might collectively refine analytical investigations based on such software applications as *Wordsmith*'s keyword function, which compare a research corpus with a significantly larger reference corpus, we would like to invite responses from colleagues working in similar areas - particularly with respect to use of the BNC or similar closed yet readily available corpora - such that we might engage in a constructive, efficiency-enhancing dialogue.

References

Aitchison, J. (1997) *The Language Web: The Power and Problem of Words.* Cambridge: Cambridge University Press.

Aitchison, J. (2001) Misunderstandings about language: A historical overview. *Journal of Sociolinguistics* **5(4)**. 611-619.

Antos, G. (1996) *Laien-Linguistik. Studien zu Sprach- und Kommunikationsproblemen im Alltag. Am Beispiel von Sprachratgebern und Kommunikationstrainings*. Tübingen: Niemeyer.

Baker, P. (2002) *Polari: The Lost Language of Gay Men*. London: Routledge.

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

Baker, P., Hardie, A. & McEnery, T. (2006) *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Bauer, L. & Trudgill, P. (eds.) (1998) *Language Myths*. London: Penguin.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999) *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education Limited.

Blommaert, J. (ed.) (1999) *Language Ideological Debates*. Berlin: Mouton de Gruyter.

Bygate, M. (2004) Some current trends in applied linguistics: towards a generic view. *AILA Review* **17**. 6-22.

Cameron, D. (1995) *Verbal Hygiene*. London: Routledge.

Cameron, D. (2000) *Good to Talk? Living and Working in a Communication Culture*. London: Sage.

Cyr, D. (1999) Metalanguage awareness: A matter of scientific ethics. *Journal of Sociolinguistics* **3(2)**. 283-286.

DiGiacomo, S.M. (1999) Language ideological debates in an Olympic city: Barcelona 1992-1996. In Blommaert, J. (ed.) pp. 105-142.

Enkvist, N.E. (1964) On defining style. In Enkvist, N.E., Spencer, J. & Gregory, M. (eds.) pp. 1-56.

Enkvist, N.E. (1973) *Linguistic Stylistics*. Berlin: Mouton.

Ensslin, A. & Johnson, S. (forthcoming) Language in the news: investigations into representations of 'Englishness' using WordSmith Tools. *Corpora: Corpus-based Language Learning, Language Processing and Linguistics* **1(2)**.

Fairclough, N. (2000) *New Labour, New Language?* London: Routledge.

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

Gal, S. & Woolard, K. (2001) *Languages and Publics: The Making of Authority*. Manchester: St. Jerome.

Garrett, P. (2001) Language attitudes and sociolinguistics. *Journal of Sociolinguistics* **5(4)**. 626-631.

Heller, M. (1999a) Sociolinguistics and public debate. *Journal of Sociolinguistics* **3(2)**. 260-288.

Heller, M. (1999b) Heated language in a cold climate. In Blommaert, J. (ed.) pp. 143-170.

Hunston, S. (2002) *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Jaworski, A., Coupland, N. & Galansiński, D. (eds.) (2004) *Metalanguage: Social and Ideological Perspectives*. Berlin: Mouton de Gruyter.

Johnson, S. (2001) Who's misunderstanding whom? Sociolinguistics, public debate and the media. *Journal of Sociolinguistics* **5(4)**. 591-610.

Johnson, S., Culpeper, J.& Suhr, S. (2003) From 'politically correct councillors' to 'Blairite nonsense': discourses of 'Political Correctness' in three British newspapers. *Discourse and Society* **14(1)**. 29-47.

Johnson, S. & Ensslin, A. (forthcoming) "But her language skills shifted the family dynamics dramatically." Language, gender and the construction of publics in two British newspapers. *Gender and Language*.

Kemppanen, H. (2004) Keywords and ideology in translated history texts: a corpus-based analysis. *Across Languages and Cultures* **5(1)**. 89-106.

Kroskrity, P. (ed.) (2000) *Regimes of Language: Ideologies, Polities and Identities*. Santa Fe: School of American Research Press.

Ku, P. & Yang, A. (1999) An analysis of key words in tourism English. In Chen, Y. (ed.) pp. 232-241.

Laforest, M. (1999) Can a sociolinguist venture outside the university? *Journal of Sociolinguistics* **3(2)**. 276-281.

Lippi-Green, R. (1997) *English with an Accent: Language, Ideology and Discrimination in the United States*. London: Routledge.

McLachlan, G.L. & Reid, I. (1994) *Framing and Interpretation*. Carlton, Vic.: Melbourne University Press.

Milroy, J. (2001) Response to Sally Johnson: misunderstanding language? *Journal of Sociolinguistics* **5(4)**. 620-625.

Niedzielski, N.A. & Preston, D.R. (2000) *Folk Linguistics*. Berlin: Mouton de Gruyter.

Phillips, M. (1989) *Lexical Structure of Text*. Birmingham: University of Birmingham Press.

Rickford, J. (1999) The Ebonics controversy in my backyard: a sociolinguist's experiences and reflections. *Journal of Sociolinguistics* **3(2)**. 267-275.

Schieffelin, B.B., Woolard, K.A. & Kroskrity, P. (eds.) (1998). *Language Ideologies: Practice and Theory*. New York: Oxford University Press.

Scott, M. (1997) PC analysis of key words – and key key words. *System* **25(2)**. 233-245.

Scott, M. (1998) Focusing on the text and its key words. In Stephens, C. (ed.) pp. 152-164.

Scott, M. (2000) Focusing on the text and its key words. In Burnard, L. & McEnery, T. (eds.) pp. 103-122.

Scott, M. (2001a) Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suite of computer programs. In Ghadessy, M., Henry, A. & Roseberry, R. (eds.) pp. 47-67.

Scott, M. (2001b) Mapping key words to *problem* and *solution*. In Scott, M.& Thompson, G. (eds.) pp. 109-128.

Scott, M. (2002) Picturing the key words of a very large corpus and their lexical upshots or getting at the Guardians' view of the world. In Kettemann, B.& Marko, G. (eds.) pp. 43-50.

Scott, M. (2004) *WordSmith Tools. Version 4.0*. Oxford: Oxford University Press. Available online from http://www.lexically.net/wordsmith/version4/index.html [Accessed 2006-07-31].

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

Sinclair, J. (2004) Corpus and text – basic principles. Available online from http://ahds.ac.uk/linguistic-corpora/ [Accessed 2006-07-31].

Stubbs, M. (1996) *Text and Corpus Analysis. Computer-assisted Studies of Language and Culture*. Oxford: Blackwell.

Tribble, C. (2000) Genres, keywords, teaching: towards a pedagogic account of the language of project proposals. In Burnard, L.& McEnery, T. (eds.) pp. 75-90.

Williams, R. (1976) *Keywords: A Vocabulary of Culture and Society*. London: Fontana.

Wolfram, W. (1998) Scrutinizing linguistic gratuity: issues from the field. *Journal of Sociolinguistics* **2(2)**. 271-279.

Sally Johnson
Department of Linguistics and Phonetics
University of Leeds
LS2 9JT

s.a.johnson@leeds.ac.uk

Astrid Ensslin
Department of Languages, Linguistics and Cultures
University of Manchester
Oxford Road
M13 9PL

astrid.ensslin@manchester.ac.uk