WORD FREQUENCY AND KEYWORD EXTRACTION

AHRC ICT Methods Network Expert Seminar on Linguistics
8 September 2006, Lancaster University, UK

‘The question is, how cruel is it?’ Keywords, Foxhunting and the House of Commons.

Paul Baker, *Lancaster University, UK.*

Keywords
fox-hunting, House of Commons, keywords, keyness, corpus, corpora, concordance, discourses, semantic, debate, saliency, frequency, aboutness.

Abstract
A small corpus of 130,000 words consisting of debates on fox hunting which took place in the British House of Commons in 2002 and 2003 was collected and then subjected to a keywords analysis. The corpus was split into two sub-corpora depending on whether speakers argued for or against fox hunting to be banned. The sub-corpora were compared together, resulting in separate keyword lists for each. Proper nouns and words relating to the debate's context (parliament) were removed from the lists prior to analysis.

This paper examines a number of keywords in detail, using concordance analyses, in order to identify different discourses (ways of looking at the world) that speakers access in order to persuade others of their point of view.

I also explore additional ways of using keyness to find salient language differences in texts, for example, by looking at key clusters and key semantic categories as well as comparing the whole corpus to a reference corpus of general British English.

Introduction
In the UK, fox hunting as it is recognised today had been practised since the seventeenth century (Scruton 1998). There have been numerous attempts to regulate or ban it, stretching back over half a century. In January 2001, according to the BBC, more than 200,000 people took part in fox hunting in the UK, and it was described as ‘one of the most divisive issues among the population’.[1] Tony Blair’s Labour Party manifesto in 1997 promised a ‘free vote in parliament on whether hunting with hounds should be banned’. In July 1999 he announced that he would make fox hunting illegal and before the next general election if possible. After a number of parliamentary debates and votes, the ban was implemented in February 2005.

In order to examine discourses surrounding the issue of banning fox hunting I decided to build a corpus of parliamentary debates on the subject. I collected electronic transcripts of three debates in the House of Commons which occurred prior to votes on hunting. These occurred on 18 March 2002, 16 December 2002 and 30 June 2003 (the total corpus size was 129,798 words). In general, the majority of Commons members voted for the ban to be ratified, although in each debate a range of options could be debated and subsequently voted upon. For example: a complete ban vs. hunting with some form of supervision.

Thinking about comparative possibilities of the fox hunting debate, it might be useful to consider that the debate has two sides and ultimately each speaker had to vote on the issue of banning fox hunting. While it may have been the case that speakers who voted the same way actually approached the subject from very different perspectives and had different reasons for the way they

voted, the fact that speakers voted, and that their contributions to the debate would be made with an idea of persuading others to vote the same way as them, suggests one area where conflicting discourses may be illuminated. Therefore, it was decided to split the corpus into two. The speech of all of the people who voted to ban fox hunting was placed into one file, while the speech of those who voted for hunting to remain was placed in another: the anti-hunt voters contributed more speech to the debates overall (71,468 words vs. 58,330 words).

Keywords in the Corpus

Using WordSmith, it is possible to compare the frequencies in one wordlist against another in order to determine which words occur *statistically* more often in wordlist A when compared with wordlist B and vice versa. Then all of the words that occur more often than expected in one file when compared to another are compiled together into another list, called a keyword list. It is this keyword list which is likely to be more useful in suggesting lexical items that could warrant further examination. A keyword list therefore gives a measure of *saliency*, whereas a simple word list only provides *frequency*.

Figure 1. Keywords when p < 0.000001.

N	WORD	FREQ.	AYE.LST %	FREQ.	XNO.LST %	KEYNESS	P
1	MICHAEL	133	0.19	0		158.8	0.000000
2	ALUN	129	0.18	0		154.1	0.000000
3	I	1,449	2.03	863	1.48	55.9	0.000000
4	CLAUSE	215	0.30	71	0.12	49.7	0.000000
5	BILL	445	0.62	220	0.38	39.0	0.000000
6	COMMONS	46	0.06	4		33.4	0.000000
7	NEW	235	0.33	99	0.17	32.9	0.000000
8	HOUSE	329	0.46	161	0.28	29.8	0.000000
9	CONCLUSION	45	0.06	5		29.2	0.000000
10	ISSUE	185	0.26	75	0.13	28.4	0.000000
11	DOGS	182	0.25	74	0.13	27.8	0.000000
12	CLEAR	112	0.16	38	0.07	24.7	0.000001
13	ATKINSON	0		15	0.03	24.0	0.000001
14	FISH	2		23	0.04	25.2	0.000001
15	MINISTER'S	10	0.01	41	0.07	27.1	0.000000
16	LIDINGTON	2		25	0.04	28.1	0.000000
17	GARNIER	3		28	0.05	28.7	0.000000
18	PEOPLE	151	0.21	222	0.38	32.0	0.000000
19	MR	316	0.44	395	0.68	32.3	0.000000
20	EXMOOR	5		38	0.07	35.9	0.000000
21	CRIMINAL	2		38	0.07	47.3	0.000000
22	GRAY	8	0.01	64	0.11	61.7	0.000000

In Figure 1, the first column (N) simply numbers the keywords in the order that they are presented (they are ordered here in terms of keyword strength). The second column (WORD) lists each keyword. The third column (FREQ.) gives the frequencies of each keyword as it occurred in the anti fox-hunting sub-corpus. The fourth column (AYE.LST %) shows this figure as a percentage of the whole sub-corpus. Where there is no figure at all, it is because the percentage is so small to be negligible. The fifth and sixth columns show the same figures for the pro-fox-hunting sub-corpus. Due to the fact that the two sub-corpora are of different sizes, the best way to compare frequencies is to look at the percentage columns rather than the raw frequency columns. The seventh column assigns a keyness value to each word; the higher the score, the stronger the keyness of that word, whereas

the final column gives the p value of each word. As p is set so low here, almost all of the figures in this column are 0.000000. Therefore the keyness value gives a more gradable account of the strength of each word in the table.

In Figure 1, the keyness score starts high (at 158.8) for the word *Michael*, and gradually decreases, to around 24 by the middle of the table. However, after that it starts to get higher again. By the last row of the table it has risen to 61.7. This is because the table is actually showing two sets of keywords (hence the fact that about half of the list is in a different colour to the other). The first part shows words which occur more frequently in the anti-hunt speeches when compared to the pro-hunt speeches, while the opposite is true for the second part of the list.

Analysis of Keywords

The majority of the keywords found consist of what Scott (1999) calls the 'aboutness' variety (words that tell us about the genre of the corpus), in both parts of the list. It should be noted again that the words at the extremes of the keyword list are the strongest in terms of them occurring *significantly* more often in one side of the debate than the other. Consider the word at row 21 of the table – *criminal*. If the proper noun *Gray* is discounted, the word *criminal* is the strongest keyword used by those who were opposed to a ban on hunting. It occurs 38 times in the collective speech of the pro-hunters and only twice in the speech of the anti-hunters. Why is this the case? As with ordinary frequency lists, this is unfortunately where the limitations of keyword lists come into play. We may want to theorise for the reasons why *criminal* is used so much by pro-hunters – looking at some of the other keywords may provide clues. However, without knowing more about the context of the word *criminal*, as it is used in both sides of the debate, our theories will remain just that – theories. Therefore, it is necessary to examine individual keywords in more detail, by carrying out concordances of them.

When a concordance of *criminal* was carried out on the corpus data (see Table 1 for an excerpt of this concordance), it was found that common phrases containing the word *criminal* included *the criminal law* (14), *a criminal offence* (10), *criminal sanctions* (6) and *a criminal act* (3). The modal verbs *would* and *should* occur as strong collocates of criminal, as do forms of the verb MAKE (e.g. *make* and *made*).

What seems clear from the table is that the pro-hunters are using a strategy of framing the proposed fox-hunting ban as criminalizing people and that they are against this. For example, the use of INVOKE in lines 2 and 4 and IMPOSE in lines 9 and 10. Here again, in order to get a better idea of the discourse prosodies associated with these terms, it is useful to refer to a corpus of general English.

Table 1. Concordance of *criminal*.

1	ack Benches. The Bill will turn into a	criminal	offence an activity now lawfully enjoyed by a
2	be particularly wrong to invoke the	criminal	law against people in my constituency who t
3	be found so to do. It is the use of the	criminal	law that would most appal me. I shall not hav
4	eman to say that the invocation of the	criminal	law in these circumstances is somehow akin t
5	Mr. Garnier: We are extending the	criminal	law. Does my hon. Friend think it in the least
6	he reason we do not normally use the	criminal	law in areas of this kind. Of course, we use t
7	sued by the new authority would be a	criminal	act attracting a fine of up to £5,000. The auth
8	his view, it should not be part of the	criminal	law. My hon. Friend the Member for North
9	ny law that we might pass. Imposing	criminal	sanctions on anybody is a serious matter. The
10	like to address the issue of imposing	criminal	sanctions on people who transgress any law t

Interestingly, in the British National Corpus (a reference corpus of 100 million words of written and spoken general British English), INVOKE collocates strongly with two sets of words – legal terms (*procedure*, *jurisdiction*, *law*, *legal*) and terms relating to supernatural forces: *spirits*, *command*, *powers* and *god*. Semantically then, INVOKE implies reference to higher powers (with a connection being made between the legal and the supernatural). The lemma IMPOSE, on the other hand, collocates in the BNC with *restrictions*, *sanctions*, *curfew*, *fines*, *ban*, *penalties*, *burdens* and *limitations*. It therefore contains an extremely negative discourse prosody – if we use *impose* in relation to *criminal law*/*sanctions*, then we are showing that we disapprove of the *criminal law*/*sanctions*.

It can therefore be seen that once a keyword is made the subject of concordance and collocational inquiry, interesting patterns of discourse begin to emerge. Terms like *invoke* and *impose* are rhetorical strategies, used to strengthen a particular discourse position, in this case, that a ban on hunting would be wrong.

What of the other keywords in the list? Due to space limitations, it is not possible to examine each one in detail, although all provide something interesting – each is a different piece of a puzzle which gradually helps to form a clearer picture. The word *people*, for example, which is key in the pro-hunt side of the debate is often used in attempts to reference a large uncountable mass in two ways. First, *people* refers to those who will be adversely affected by the Bill if it is passed (their livelihoods stopped, their communities threatened and their futures involving a prison term). Secondly, it refers to (a presumably greater number of) people who do not hunt, but are not upset or concerned by those who do.

However, the keyword list has only given us a small number of words to examine, and once all of the proper nouns (*Michael*, *Alun*, *Atkinson*, *Lidington*, *Garnier*, *Gray*) have been discounted, this leaves us with just sixteen words in total. We may also want to discount (or at least background for the moment) the keywords which relate to text genre, in this case parliament (*Bill*, *Commons*, *House*, *Minister's*), which leaves us with only twelve keywords.

Twelve keywords do not give us much to analyse. So in order to address this issue, the p value was increased to $p < 0.001$ and the keywords process was carried out again, eliciting 120 keywords, which was reduced to 88 once the proper nouns were discarded.

Although the keyness scores in this longer list are less impressive, what is interesting about working with a larger list, is that it becomes possible to see connections between words, which may not always be apparent at first, but are clearer once they have been subjected to a more rigorous mode of analysis. For example, keywords in the pro-hunt debate include the following words: *fellow*, *citizens*, *Britain*, *freedom*, *imposing*, *illiberal*, *sanctions* and *offence*. All these keywords are connected in some way to the findings we have already looked at. So *sanctions*, *offence*, *imposing* and *illiberal* occur in similar ways to the word *criminal* which was examined above.

As a different yet related strategy, the keywords *fellow*, *citizens*, *Britain* and *freedom* are related to the keyword *people* which was discussed earlier. Consider the concordance in Table 2. We can see that the term *fellow citizens* is always preceded by a first person possessive pronoun (*my* or *our*). The use of this term looks like a strategy on the behalf of pro-hunters to appear to be speaking for and with the people of Britain, thereby implicitly labelling their discourse as a hegemonic one. Note also how in lines 10 and 11, the debater actually speaks for the people 'the people of Britain are beginning to catch on', 'for most of the 55 million people in England it is of peripheral interest'. Finally, in lines 13-16 the lemma RESTRICT and the word *individual* both collocate with *freedom*. There is an underlying nationalist discourse being drawn on here, which could be paraphrased as: 'Britain is a good country because it is a place where people are free'. This discourse is used as an argument to allow fox hunting to continue.

Table 2. Sample concordance of *fellow citizens*, *Britain* and *people* (pro-hunt).

1	able to me and, I believe, to most of my	fellow citizens	. The killing of an animal is justifiable only
2	a small but significant minority of our	fellow citizens	. I agree with one thing the Minister said.
3	al freedom, that it will rob some of our	fellow citizens	of their livelihood and take homes from a
4	7, when the pensions of millions of our	fellow citizens	are affected by a deeply serious crisis fr
5	at. Of course, I accept that some of our	fellow citizens	genuinely disapprove of hunting with hou
6	umber of my family and 407,000 of my	fellow citizens	, I took part in the march for liberty and liv
7	the Third Reich. Down the ages, we in	Britain	have fought against the persecution of min
8	an who ripped apart the fabric of rural	Britain	and passed the most illiberal and divisive p
9	that is being practised on the people of	Britain	tonight. Mr. Atkinson: There we have it.
10	se to offer the people of Britain, and the	people	of Britain are beginning to catch on.
11	rs speak, but for most of the 55 million	people	in England it is of peripheral interest. Mr.
12	ce to a largely urban nation, millions of	people	people recognise that to criminalise at a str
13	unjustifiable restrictions on individual	freedom	, would increase the suffering of foxes
14	unjustifiable restrictions on individual	freedom	, that it will rob some of our fellow citiz
15	t, illiberal and arbitrary. It will restrict	freedom	and do nothing to help animal welfare.
16	e unjustifiable restrictions on individual	freedom	trying to justify itself, but failing, in th

Therefore, examining these additional keywords helps to build on the findings we have already uncovered. A number of discourses are then starting to come into focus, particularly for the pro-hunt speakers at this stage. For example, use of terms like *criminal*, *sanctions*, *offence* and *imposing* suggest a discourse of civil liberties, whereas words like *Britain*, *fellow*, *citizens* and *people* suggest a discourse of shared British identity.

Using a Reference Corpus

So far our keywords analysis has been based on the idea that there are two sides to the debate, and that by comparing one side against another we are likely to find a list of keywords which will then act as signposts to the underlying discourses within the debate on fox hunting. Our analysis so far has uncovered some interesting differences between the two sides of the debate. However, it also raises some issues. In focussing on difference, we may be overlooking similarities – which could be equally important in building up a view of discourse within text. For example, why do certain words *not* appear as keywords? Considering that *barbaric* occurred as a keyword in the anti-hunting speeches, another word that I had expected to appear as key in the anti-hunting debates was *cruelty*. However, this word occurred 124 times in the anti-hunting speeches and 106 times in the pro-hunting speeches. In terms of proportions, taking into account the relative sizes of the two sub-corpora, the anti-hunt speakers actually used the word *cruelty* proportionally *less* than the pro-hunters (0.17% vs. 0.18%). So while *cruelty* occurred slightly more often on one side of the debate, this was not a statistically significant difference – clearly the concept of cruelty is important to both sides. However, how would we know (without making an educated guess) that a word like *cruelty* is worth examining? One solution would be to carry out a different sort of keywords procedure; this time by comparing the entire set of debates against another corpus – one which is representative of general language use. This would produce a keyword list that highlights all of the words which occur in the fox hunting debates more frequently than we would expect in 'normal' language. In this case it was decided to implement the Freiberg-Lancaster/ Oslo-Bergen (FLOB) corpus which consists of one million words of written British English taken from the 1990s. Although the FLOB corpus contains written texts and the debates were spoken, a good proportion of the debate consists of prepared speech, so in a sense it could be argued as having elements of written language within it.

A comparison of the hunting debates with the FLOB corpus reveals a different set of keywords; the twenty strongest being *hon.*, *hunting*, *that*, *bill*, *ban*, *I*, *friend*, *Mr*, *foxes*, *member*, *clause*, *fox*, *minister*, *cruelty*, *we*, *gentleman*, *house*, *my*, *dogs* and *is*.

Comparing this list to Figure 1 (which showed keywords when the two sides of the debate were examined), it is clear that some of these words are key in the debates when compared to FLOB because they occur very frequently on one side of the debate (for example, *I*, *clause*, *bill*, *house* and *dogs* are key in both lists due to their prevalence of use by anti-hunting speakers). However, other words do not appear in both lists, for example *foxes* and *cruelty*. A further line of investigation therefore could be to examine words which are key across the debate when compared to a reference corpus, rather than simply looking at words which are only key on one side of the debate.

Examining the word *cruelty* in more detail, it becomes apparent that although it occurs with a reasonably comparable frequency on each side of the debate, the ways that it occurs are quite different for different speakers. The anti-hunters tend to use it in conjunction with words like *ban*, *outlaw*, *unnecessary*, *target* and *eradicate* (Table 3). Their speech also tends to assume that cruelty already exists e.g. 'The underlying purpose of the Bill is to ban all cruelty associated with hunting with dogs.' However, those who are pro-hunting question this position – using collocates such as *test*, *tests*, *prove*, *evidence* and *defining* (Table 4). Therefore rather than accepting the presence of cruelty, pro-hunting speakers problematise it: e.g. the full text in line 1 of Table 4 is: 'Cruelty is subjective and comparative, and the Bill entirely fails adequately to define cruelty or utility.'

Table 3. Concordance (sample) of cruelty (anti-hunt).

1	romise, no uncertainty, no delay; a *ban* on the	cruelty	and sport of hunting in the lifetime of this
2	ael: I see it very clearly in a Bill that *bans* the	cruelty	associated with hunting in all its forms. I h
3	ate about banning cruelty and *eradicating* the	cruelty	associated with hunting. I have tried to be
4	law, to be enforceable and to *eradicate* all the	cruelty	associated with hunting with dogs, and I i
5	rtant issue for many who want to see an *end* to	cruelty	and for those who want things to remain a
6	listen to an organisation that exists to *prevent*	cruelty	to animals and I remind the hon. Member
7	enshrining in law the principle of *preventing*	cruelty	as well as the principle of recognising utili
8	ke effective and enforceable law. It will *tackle*	cruelty	, but it also recognises the need to deal wi
9	ise, is uncompromising in seeking to *root out*	cruelty	. It will not allow cruelty through hunting
10	mingly, twice, to bring an end to *unnecessary*	cruelty	to wild mammals. There can seldom in pa

Table 4. Concordance (sample) of *cruelty* (pro-hunt).

1	and the Bill entirely fails adequately to *define*	cruelty	or utility. As my hon. Friend the Mem
2	eedless or avoidable suffering" when *defining*	cruelty	. The phrase playing the fish" is no euph
3	al act. The arbitrary application of the *tests* of	cruelty	and utility to fohunting is illogical when
4	ul unless those who hunt can meet the *tests* of	cruelty	and utility described by the Minister. Th
5	. The whole House has heard the *definition* of	cruelty	, as given by the Minister, relating to ne
6	ften than not, focuses on cruelty or *perceived*	cruelty	. I commend the former Home Secretary
7	. It will not be for the authorities to *prove* that	cruelty	takes place; if the Bill is enacted, hunti
8	r described as incontrovertible *evidence* of the	cruelty	of deer hunting, he must tell us what it i
9	s. If the Minister is so concerned, where is the	cruelty	*test* in the autumn for shooting or snari
10	he Minister said that those would not pass the	cruelty	or utility *tests*. How can he know that?

Comparing a smaller corpus or set of texts to a larger reference corpus, is therefore a useful way of determining key concepts across the smaller corpus as a whole. Indeed, for many studies where the text or set of texts under scrutiny is relatively uniform, using a reference corpus may be all that is needed. However, in order to address the problem of over-focussing on differences at the expense of similarities, it is recommended that the corpus being analysed is used in the creation of more than one keyword list.

Key Categories

A further way of considering keyness is to look beyond the lexical or phrasal level, for example by considering words that share a related semantic meaning or grammatical function. While a simple keyword list will reveal differences between sets of texts or corpora, it is sometimes the case that lower frequency words will not appear in the list, simply because they do not occur often enough to make a sufficient impact. This may be a problem, as low frequency synonyms tend to be overlooked in a keyword analysis. However, text producers may sometimes try to avoid repetition by using alternatives to a word, so it could be the case that it is not a word itself which is particularly important, but the general meaning or sense that it refers to. For example, the notion of 'largeness' could be key in one text when compared to another, and this would be demonstrated by the writer using a range of words such as *big*, *huge*, *large*, *great*, *giant*, *massive* etc – none of which occur in great numbers, but taken as a cumulative whole, would actually appear as key. Thinking grammatically, in a similar way, one text may have more than its fair share of modal verbs or gradable adjectives or first person pronouns when compared to another text. Finding these *key categories* could help to point to the existence of particular discourse types – they would be a useful way of revealing discourse prosodies.

In order for such analyses to be carried out, it is necessary to undertake the appropriate form(s) of annotation. The automatic semantic annotation system used to tag the fox hunting corpus was the USAS (UCREL Semantic Analysis System) (Wilson and Thomas 1997). This semantic tagset was originally loosely based on McArthur's (1981) *Longman Lexicon of Contemporary English*. Once the semantic annotation had been carried out, word lists (consisting of words and semantic tags) of the two sides of the fox hunting debate were created and compared with each other to create a keyword list. From this list, the relevant key semantic tags were singled out for analysis. There isn't enough space to look at all of the key tags in detail, so I want to concentrate on a couple of significant findings here.

Two key tags which occurred significantly more often in the pro-hunt speeches were S1.2.6 'sensible' and G2.2 'ethics – general'. Looking at a concordance of words that were tagged as S1.2.6 (Table 5 shows a small sample from of the total number of cases) it is clear that this contains a list of words relating to issues of sense: *sensible*, *reasonable*, *common sense*, *rational*, *ridiculous*, *illogical*

and *absurd*. The prevalence of this class of words is due to the way that the pro-hunt speakers construct the proposed ban on hunting (as ridiculous, illogical and absurd) and the alternative decision to keep hunting (as reasonable, sensible and rational). While this way of presenting a position would appear to make sense in any argument it should be noted that the anti-hunt speakers did *not* tend to characterise the debate in this way. They did not argue, for example, that their position was sensible, reasonable etc. and that of their opponents was ridiculous and absurd. It is also worth noting that one feature of hegemonic discourses is that they are seen as 'common-sense' ways of thinking. To continually refer to your arguments in terms of 'common-sense' is therefore a powerful rhetorical strategy. With this sort of analysis, we are not only seeing the presence of discourses in texts, but we are also uncovering evidence of how they are repeatedly presented as the 'right' way of viewing the world.

Table 5. Concordance (sample) of words tagged as S1.2.6 'sensible' (pro-hunt).

1	he Bill makes illegal only the perfectly	reasonable	sensible and respectable occupations
2	continuation of hunting. I appeal to all	reasonable	hon. Members to support me in seeki
3	inal law rather than fiddle around in an	absurd	way with this absurd Minister on this
4	rmed roast. The debate has not shown a	rational	analysis of the facts: misplaced co
5	be justified by scientific evidence. The	ridiculous	new clause 13 wrecks it further, and i
6	this matter. Most people with common	sense	will say, "Why don't they reach a dea
7	eds your protection. Mr. Gray: Calm,	sensible	and rational people across Britain a
8	ss. Why not? That would be a logical,	sensible	and coherent approach. As I have to
9	method of control in that time is utterly	illogical	Mr. Gray: My hon. Friend makes an
10	ng-during that time. This ludicrous and	illogical	new clause is the result of a shabby d

What other key categories of meaning did the pro-hunters tend to focus on? The G2.2 tag was affixed to a set of words relating to ethics, including *moral*, *rights*, *principles*, *humane*, *morality*, *ethical*, *legitimate*, *noble* and *fair*. It appears that the pro-hunt speakers are more likely to argue their position from an explicitly ethical standpoint – a somewhat surprising finding considering that the ethical position of ending cruelty to animals would appear to be a more obvious stance for the anti-hunt protesters to have taken. However, a closer examination of a concordance of words which receive the G2.2 tag (Table 6) reveals that the pro-hunt speakers are pre-occupied with issues of morality because they wish to question the supposed absolutist ethical standpoint of the anti-hunters. Therefore, their frequent references to ethics are based around attempts to problematize or complicate the ethical position of the anti-hunters: again, this finding complements and widens the analysis of the word *cruelty* above.

Table 6. Concordance (sample) of words tagged as G2.2 – 'ethics: general' (pro-hunt).

1	e should be careful about imposing our	morality	on other people, someone on the Lab
2	ople to make up their own minds about	morality	. One of the issues that I dealt with as
3	In any event, they are surely moral and	ethical	issues to be considered by individu
4	g, vivisection and slaughter? There are	moral	gradations here and no moral absolut
5	the Bill that it is based on no consistent	ethical	principle. I was rather pleased when
6	ere is a complete absence of consistent	ethical	principles in the contents of the Bill.
7	at not an issue? Is hunting not the more	humane	method of controlling the fox pop
8	omeryshire (Lembit Öpik). There is no	moral	justification for the Government's po
9	questions involved, will he explain the	moral	difference between a gamekeeper us
10	en. Predators do not consider the moral	rights	and wrongs as we do as human bein

What about the other side of the debate? One semantic category which occurred more often in the speech of those who are opposed to hunting was S1.2.5 'Toughness; strong/weak'. This category consists of words such as *tough*, *strong*, *stronger*, *strength*, *strengthening*, *robust*, *weak* and *feeble* (Table 7 shows a small sample of these cases). On this side of the debate then, the pro-hunt stance is viewed as weak, whereas the proposed Bill is frequently characterised as tough, strong or robust. So here we have a significant difference in the ways that the two sides of the debate try to position themselves as correct. While the pro-hunt debate frames itself in terms of what is *sensible*, the anti-hunt debate uses *strength* as its criteria.

Table 7. Concordance (sample) of words tagged as S1.2.5 5 'Toughness; strong/weak' (anti-hunt).

1	to the Bill, we would have incredibly	strong	legislation with which to tackle hunti
2	lleagues to unite today in getting good,	strong	legislation through the House. I hope
3	n. However, although the current Bill is	strong	in that respect, it does not set the th
4	hon. Lady's argument is not especially	strong	. The Bill is good in that it takes us
5	stands is far from imperfect. It is a very	strong	Bill. It deals with the issue of cruelty
6	the other Government amendments to	strengthen	the Bill are agreed, I can give the Ho
7	practicable in their area. The measure is	tough	but fair, and it will be simple to
8	The tests, as I have said, are	tough	but fair. Supporters of hunting say th
9	eve in while being seen by the public as	tough	and fair and being strong enough to
10	upport it appear to be unable to see the	weakness	of their case. Having given every op

A semantic tagging of the corpus then, helps to reveal some of the more general categories of meaning which are used in the construction of discourse positions on the different sides of the debate. The pro-hunt speakers talk in terms of what is sensible, whereas the anti-hunt speakers talk in terms of what is strong. On their own, individual words like *strong*, *tough*, *sensible* and *rational* did not appear as keywords – it was only by considering them as a single part of a wider semantic category that their importance became apparent. Widening the scope of keywords beyond the lexical level can therefore be a fruitful endeavour.

Conclusion

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

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

References

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

Scott, M. 'Picturing the Key Words of a Very Large Corpus and their Lexical Upshots – or Getting at the Guardian's View of the World' in B. Kettemann and G. Marko (eds.) *Teaching and Learning by Doing Corpus Analysis* (Amsterdam: Rodopi, 2002), 43–50.

Scruton, R., *On Hunting* (London: Yellow Jersey Press, 1998).

Wilson, A. and Thomas, J. 'Semantic Annotation', in R. Garside, G. Leech and A. McEnery (eds) *Corpus Annotation: Linguistic Information from Computer Texts* (London: Longman, 1997) 55–65.

Notes

[1] http://news.bbc.co.uk/1/hi/uk/449139.stm. (date accessed 11 May 2006)