O'Keeffe, A., (2012) Corpora and Media Studies. In K. Hyland, M. H. Chau and M. Handford (Eds) Corpus Applications in Applied Linguistics. London: Continuum, pp. 117-131.

Corpora and Media Studies

Anne O'Keeffe

Introduction: getting started

Traditionally, studies in media discourse have been divided into those that focus on spoken media (mostly radio genres) and those that focus on written media (mostly newspapers). Studies into spoken media discourse were largely covered by conversation analysts (Conversation Analysis, see Hutchby 1991, 1996) and written media discourse was more likely to be explored within a critical framework (Critical Discourse Analysis, see Fairclough 1995a, 1995b, 2000). Considering the prevalence of media discourse in everyday life, the number of studies based on it as a whole over the years is less than one would expect. Reasons for this probably lie in the difficulty of gathering data. In the case of spoken data, it has to be recorded and transcribed, a time-consuming and laborious task. In the case of written discourse, previous to the advent of the internet, the data needed to be scanned (and checked) or keyed into a computer. It is not a coincidence therefore that most studies of media discourse up until the year 2000 or so, whether spoken or written, focused on small amounts of data that did not need much recording, transcription or scanning time.

For corpus linguists, accessibility of data is always a key issue and despite all of the technological advances, the drudge of transcription has not really gone away in relation to the assembling of spoken media data into a corpus. It is relatively easy now to build a corpus of newspapers. With interfaces such as *Lexis-Nexis*, one can quite readily assemble a large research corpus (obviously within the bounds of any copyright restrictions that may prevail). Even on a small-scale, newspaper stories can generally be gathered quickly online. When it comes to the spoken word however, while it is infinitely easier to get access to radio or television material online, the scourge of transcription still prevails. This has inhibited the growth of studies in relation to spoken media genres. Despite this, there are a number of ways in which spoken data can be gathered online for research purposes. Some news corporations provide transcripts of political interviews and

many avid fans painstakingly transcribe interviews with their hero and post them online. For example, if you enter an internet search query for the famous name plus the word "interview", you will usually get a number of sites where interviews have been transcribed. These will, almost without exception, have been cleaned up in terms of real-time discourse features such as hesitations, false starts and repetitions but generally, they will not have been altered beyond recognition. It is usually easy to access a recording of the actual interview and so one can then add back in the features that have been cleaned up and check the authenticity of the transcript.

Obviously, in building a corpus of media discourse, the same principles prevail in relation to representativeness. A corpus needs to have a principled underpinning. A collection of texts is not necessarily a corpus which can produce reliable results. It has to be designed around a solid design matrix. In the case of newspaper studies, the corpus needs to have parameters of time (from X date/year to Y date/year – with a rationale as to why these dates were chosen). The newspaper corpus needs to consider newspaper type (broadsheet versus tabloid, national versus regional,and so on). Other considerations include, whether the corpus contain only news reports, or include editorials. Will it focus on one particular topic? Many considerations need to be taken into account.

In relation to spoken data, there are different considerations. In trying to represent spoken media interactions, O'Keeffe (2006) put together a small corpus that was centred around personae type, that is the data was collected on the basis of trying to represent the voices and identities with the interactions of:

1) political persona (people in the political sphere, radio and television political interviews made up this sub-category)

2) public persona (people who were in the political sphere but who were well-known in the public sphere, television chatshows formed a major part of this sub-category)

3) private persona (people who phoned radio programmes but who were not known in the public sphere – they were using their private sphere identities, radio phone-ins made up most of this sub-category).

This allowed for the gathering of data around three different interaction types. Not all media discourse is based on political or news interviews, not all is television chat show nor is it radio phone-in. All needed to be represented in equal measure.

What can a corpus do for the study of media discourse?

The core functions of corpus software, namely the generation of key word and word frequency lists and concordances, can offer much to the analysis of a corpus of spoken or written media discourse. In this section, we will survey their main applications and use concrete examples from media interviews.

Keywords

Keywords are not necessarily the most frequent words in your corpus rather they are the most unusually frequent words. It is very easy to calculate keywords using corpus software such as *Wordsmith Tools* (Scott 2008). Essentially, the software compares the word frequency list of the text or corpus which you are focusing on with some other larger 'reference corpus'. The choice of reference corpus can have an influence on the results. For example, if we take an internationally recognisable interview and do keyword calculations against different reference corpora, you will see the varying results. The interview that we will focus on is the *BBC 1 Panorama* television interview by Martin

Bashir of Diana, Princess of Wales (broadcast November 1995). The transcript is readily available on the internet as is the actual television interview, see: http://www.bbc.co.uk/politics97/diana/panorama.html.

First, we compare this interview with a corpus of media interviews from O'Keeffe (2006) which is made up of 271,553 words from 29 political interviews (93,180), 46 interviews on TV chat shows and radio involving known or public persona (89,225) and 17 interviews from radio phone-ins involving unknown of private persona (89,148). Data is drawn from international English-speaking media sources including from the UK, USA, Canada, Australia and Ireland.

The key words in this case are relatively short, 27 in all. Predictably, the names of the interviewer and the interviewee are the top items. These are in fact marking the speakers in the interview. In each table, these can be ignored for the most part.

Table 1 - Keywords of Bashir-Diana *Panorama* interview with Media corpus (O'Keeffe 2006)

1	Diana	7	marriage	13	difficult	19	were	25	media
2	Bashir	8	husband	14	William	20	yourself	26	depression
3	did	9	had	15	royal	21	because	27	husband's
4	was	10	uh	16	my	22	relationship		
5	Wales	11	monarchy	17	role	23	your		
6	prince	12	bulimia	18	queen	24	children		

When we generate the key words of the same interview against a different reference corpus, we get more and different results. This time, we use a reference corpus which is very unrelated to media interviews, namely an academic corpus of English, the Limerick-

O'Keeffe, A., (2012) Corpora and Media Studies. In K. Hyland, M. H. Chau and M. Handford (Eds) Corpus Applications in Applied Linguistics. London: Continuum, pp. 117-131.

Belfast corpus of Academic Spoken English (LIBEL). This is made up of lectures, tutorials, seminars and presentations. In this case, 500,000 words were used from LIBEL.

Table 2.0 Keywords of Bashir-Diana *Panorama* interview with LIBEL (top 50 of 94)

1	Diana	11	I've	21	people's	31	husband's	41	feel
2	Bashir	12	it's	22	bulimia	32	couldn't	42	loved
3	was	13	me	23	you're	33	divorce	43	think
4	I	14	uh	24	queen	34	that's	44	never
5	don't	15	yes	25	William	35	people	45	wasn't
6	husband	16	didn't	26	were	36	difficult	46	Mr.
7	my	17	had	27	because	37	public	47	princess
8	I'm	18	prince	28	monarchy	38	there's	48	royal
9	did	19	I'd	29	myself	39	yourself	49	pressures
10	Wales	20	marriage	30	role	40	relationship	50	albeit

In all, there were a total of 94 key words from this list which were statistically 'key', that is, occurring with unusual frequency when compared with academic lectures and seminars. Table 1, the key words from the comparison with the media corpus, contained only 27 words in all. Because the academic data is more generically diametric, it results in a broader range of key items, including common first and second person pronouns *I, I'm, my, myself, yourself, me,* high frequency verbs and verb forms *was, did, didn't, wasn't loved, think,* pronoun verb combinations *I've, you're,* everyday seeming nouns *divorce, husband, marriage, people's, husband's,* and so on. Of course, these all relate to more private sphere domains of reference 'you – I', relationships, marriage, problems like bulimia, marriage breakdowns, all of which would not normally be talked about in the more referential world of academia, as these short extracts from the *Panarama* interview in question and the data from LIBEL illustrate:

1) extract from interview between Martin Bashir and Diana, Princess of Wales, broadcast November 1995

BASHIR: What effect did the depression have on your marriage?

DIANA: Well, it gave everybody a wonderful new label - Diana's unstable and Diana's mentally unbalanced. And unfortunately that seems to have stuck on and off over the years.

BASHIR: Are you saying that that label stuck within your marriage?

DIANA: I think people used it and it stuck, yes.

2) extract from a literature lecture (LIBEL)

….So according to David Lloyd's writings we tease out concepts of national identity through literature. Germany for example has a great philosophical tradition. The strange thing about Ireland is that we do not have a great philosophical tradition. We have a great literary tradition. Germany has a great philosophical tradition. The French in the nineteen –sixties and seventies produced a terrific theoretical body of work. In Ireland largely we have used literature to sculpt concepts of national identity.

Let us compare the *Panorama* Bashir-Diana interview with two further datasets. Both of these have in common with the interview the fact that they involve more reference within the 'I – you' domain and they refer more to everyday worlds of relationships, and so on. These are a corpus of the sitcom programme *Friends* (based on Malveira Orfano 2010) and secondly, the Limerick Corpus of Irish English (LCIE), a one million-word corpus of everyday spoken Irish English (see Farr et al 2004).

Table 3 - Keywords Bashir *Panorama* interview with *Friends* sitcom corpus (50 in total)

1	Diana	11	media	21	William	31	people's	41	get
2	Bashir	12	and	22	Obviously	32	bulimia	42	can
3	was	13	were	23	Royal	33	queen	43	not
4	very	14	prince	24	Country	34	effect	44	look
5	people	15	role	25	Monarchy	35	yourself	45	is
6	husband	16	difficult	26	As	36	interest	46	know

7	had	17	children	27	relationship	37	being	47	you
8	of	18	marriage	28	The	38	felt	48	just
9	public	19	because	29	In	39	think	49	no
10	Wales	20	that	30	Did	40	depression	50	oh

Table 4 - Keywords Bashir *Panorama* interview with LCIE as reference corpus (top 50 of 89 keywords)

1	Bashir	11	bulimia	21	Uh	31	feel	41	pressures
2	Diana	12	role	22	My	32	your	42	separation
3	husband	13	difficult	23	people's	33	princess	43	attention
4	marriage	14	royal	24	Very	34	was	44	engagements
5	Wales	15	William	25	Yes	35	effect	45	daunted
6	prince	16	because	26	husband's	36	book	46	were
7	relationship	17	public	27	Divorce	37	children	47	future
8	media	18	queen	28	depression	38	obviously	48	enormous
9	monarchy	19	had	29	Yourself	39	did	49	knowledge
10	people	20	that	30	Albeit	40	being	50	duties

As tables 3 and 4 illustrate, and in common with table 2, the spread of key words from the Bashir-Diana interview is broad, much broader than when you compare it with the corpus of media interviews (table 1). This tells us that if you do a key word analysis using a reference corpus that is very like the test corpus, you will get more concentrated (and fewer) key word forms. If you take a corpus which is very different to the test corpus, as in the case of LIBEL, you will get a very diverse range of key words including some that you will not expect (e.g. all of the first and second person pronouns and high frequency verbs). If you take a general corpus, representing how English is generally used (in this case LCIE), then you will still get a lot of key words but they will be less disparate. If you look at the results in tables 3 and 4, you see that they have a lot in common with tables 1 and 2 but they have a broader spread. Many of the noun forms are common to all four tables, e.g. *husband*, *bulimia* and *monarchy*. Interestingly, *divorce*, is on tables 2, 3,

and 4 but not table 1 when the key words are generated by comparison with other media interviews. This suggests that *divorce* is not an uncommon reference in media interviews.

Let us look at the other instruments to hand, namely, word frequency lists and concordance lines.

Word frequency lists

Word frequency lists can be easily generated by corpus software and they simply refer to the rank ordered frequency of all of the words, or 'types' in the whole of your corpus. It is very often illuminating to compare where in the rank order a word is in comparison to some other baselines. For example, in table 5, we put the top 20 most frequent forms in the Bashir-Diana *Panorama* interview, the media, LIBEL, *Friends* sitcom, and LCIE corpora.

Table 5 - Top 20 words in Bashir-Diana *Panorama* interview, media corpus, LIBEL, *Friends* sitcom corpus and LCIE everyday conversations sub-corpus

N	Bashir Diana	Media corpus	LIBEL	Friends	LCIE everyday conversations
1	I	the	the	I	the
2	the	and	you	you	I
3	and	I	and	the	and
4	to	to	to	to	you
5	you	a	of	a	to
6	that	of	that	and	it
7	a	you	a	it	a
8	was	that	in	that	yeah
9	it	in	it	what	that
10	of	it	I	oh	in
11	Diana	was	is	is	of
12	Bashir	is	s	no	was
13	in	on	so	okay	like
14	but	have	we	know	is
15	my	we	what	my	know
16	what	but	amm	this	he
17	do	for	this	of	on

18	did	they	there	yeah	no
19	had	this	okay	me	they
20	people	be	have	do	but

Based on this we can make a number of observations:

- Comparing the *I – you* domain on the datasets is a good starting point. In conversation, these two pronouns are usually very high in the frequency list and this is a marker of the interactive nature of conversation. In table 5, we can see that *I* and *you* are high ranking in all datasets but in the academic lecture data, we see that there is a lot more *you* reference than *I.*
- Anomalies always need to be checked out by looking at concordance lines to try to get to the bottom of why certain words have floated to the top, so to speak. In the following section, we will follow up on the follow words which are high-frequency in the Bashir-Diana interview but not to the same degree in the other datasets, these are *was, my, what, do, did, had, people.*

Concordances

Sinclair (1991: 32) tells us that 'a concordance is a collection of the occurrences of a word-form, each in its own textual environment' and that 'in its simplest form it is an index. Each word-form is indexed and a reference is given to the place of occurrence in a text.' A set of concordance lines then is in essence a list of occurrences of a search word, where all of the occurrences of that word, throughout the corpus, or text, will be collated into one place and indexed back to their position within the dataset. The search word is usually referred to as the node word and it normally appears in the centre of the concordance lines. A spin off benefit of having all of the node words aligned down the centre of the page is that the eye can pick out salient patterns that emerge to the left and right of the node. These can usually be computed by the corpus software by way of follow up but very often the researcher will form hypotheses on scanning the patterns in the concordance itself. Using the concordance function then, let us look at one of the

anomalies that we have identified from the word frequency lists in table 5, namely *was*, by way of example of the type of research process that would follow:

Was

The appearance of the past tense verb form *was* in the top 20 items in the interview begs explanation. The only other dataset where we find it on table 5 is in LCIE, i.e. casual conversation where it is used within narratives. It is worth exploring whether it is being put to a similar use in this interview. Here is an extract of the concordance lines for *was* using *Wordsmith Tools* (Scott 2009):

Figure 1 – A sample from the concordance lines for *was* from Bashir – Diana interview

7	s of questions - and I hoped I	was	able to reassure them. Bu
9	s agendas changed overnight. I	was	now separated wife of the
10	e of Wales, I	was	a problem, I was a liability (seen as), an
11	wife of the Prince of Wales, I	was	a problem, I was a liabil
12	their attitude towards me. It	was	, you know, if we are goin
13	rpose was behind it? DIANA: It	was	to make the public change
14	d more cards than I would - it	was	very much a poker game, c
16	nuisance phone calls? DIANA: I	was	reputed to have made 300
18	adulterous relationship, which	was	not true. BASHIR: Have yo
19	presses his affection for you.	Was	that transcript accurate?
20	ional press? DIANA: No, but it	was	done to harm me in a seri
21	t time I'd experienced what it	was	like to be outside the ne
22	in a serious manner, and that	was	the first time I'd experi
23	heard it on the radio, and it	was	just very, very sad. Real
24	Andrew Morton's book about you	was	published. Did you ever m
25	her `annus horribilis', and it	was	in that year that Andrew
26	saw the distress that my life	was	in, and they felt it was
27	, I did. BASHIR: Why? DIANA: I	was	at the end of my tether.
28	life was in, and they felt it	was	a supportive thing to hel
29	od team in public; albeit what	was	going on in private, we w
30	uired. BASHIR: Do you think it	was	accepted that one could l
31	mily? DIANA: I think everybody	was	very anxious because they
32	A: No, because again the media	was	very interested about our

At first there seems to be no order to this stream of seemingly truncated lines but any one line can be clicked to bring the researcher back to the source file. In addition, as referred to above, when looking at a search word in context, one needs to look at the patterns that

form to its left and right. Most concordancing software will also have the facility to compute the collocates for you. Table 6 shows the breakdown of the words that go before and after *was.*

Table 6 – plot of collocates of node word *was* using *Wordsmith Tools*

N	Word	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	WAS	1	6	2	2	0	181	0	2	2	6	1
2	I	8	6	7	2	43	0	0	4	3	2	5
3	A	5	3	6	1	0	0	29	3	3	7	7
4	IT	2	0	1	1	45	0	2	2	3	2	4
5	THE	10	3	4	12	0	0	11	4	5	6	6
6	AND	5	7	4	13	2	0	0	4	4	5	3
7	THAT	3	2	7	6	10	0	2	4	2	6	5
8	TO	2	4	5	1	0	0	3	13	8	5	5
9	OF	0	2	4	2	0	0	0	2	7	5	3
10	DIANA	0	5	10	6	1	0	0	0	1	0	0
11	IN	5	2	0	1	0	0	4	2	4	2	3
12	BECAUSE	1	1	1	5	0	0	1	3	2	8	0
13	WHAT	1	1	2	1	13	0	1	0	1	1	0
14	BUT	1	1	3	9	0	0	1	1	2	1	2
15	YOU	4	3	3	1	1	0	1	1	1	4	1
16	MY	0	2	2	4	0	0	5	1	0	3	3
17	BASHIR	2	1	1	10	1	0	0	0	0	0	0
18	VERY	0	0	0	0	0	0	9	4	1	0	0
19	YOUR	1	0	0	2	0	0	5	1	0	2	2
20	OUT	3	1	0	0	0	0	2	2	3	1	0
21	WITH	1	1	3	0	0	0	0	1	2	3	0

From table 6, we can see that:

- Pronouns *I* and *it* account for the 49% of all the words that go before *was*
- *A, the* and *very* account for 27% of all the words that come after was
- *What was* accounts for 7% of all of the patterns to the left of the node word *was.*

From each of the above respectively, we can say that:

- There is a lot of representation of how the interviewee felt and how the interviewee represents situations in the past and how the interviewee felt she was represented. She cleverly merges all of these into very emotive language centring around the first person pronoun. This in turn allows the interview to be set up as a haunting self-revelation of how hard it was to be 'me', Diana. This is telling of an interview that seeks to reveal what was like to be Diana, Princess of Wales:

```
I was a different person.
I was concerned I was a fat, chubby…
I was a problem, I was a liability…
I was a problem, fullstop.
I was absolutely devastated
But I was actually crying out
I was again unstable, sick,
I was almost an embarrassment
I was ashamed because I could
I was at the end of my tether
I was compelled to go out and do my engagements
… as far as I was concerned I was a fat,
I was constantly tired, exhausted…
I was crying out for help…
I was desperate.
I was in love with him.
I was now separated wife…
```

Equally, patterns with *it was* show a negative portrayal of what it was like to be Diana, Princess of Wales:

```
And in our private life it was obviously turbulent.
I heard it on the radio, and it was just very, very sad. R
It was already difficult, …
We struggled a bit with it, it was very difficult;
Well, there were three of us in this marriage, so it was a bit crowded.
…it was very distressing for me
…it was so cruel.
…I felt it was unfair, because I want
…it was isolating…
It was a challenge…
```

- The patterns to the right of the node word all point to noun and adjective phrases: *a*, *the* and *very*. Again negativity prevails in the portrayal of life as a Princess:

Was + noun phrases with *a*:

```
a problem
a liability
a basket-case
a fat, chubby, 20-year-old
a challenge
a lot of anxiety
a pretty dull subject
a bit of a difficult time
```

Was + noun phrases with *the*

Of the 11 instances of *was* followed immediately by *the*, we see from figure 2, four are of interrogative forms in questions from Bashir (*what was…?;* and *Who was…?*). We will return to the interrogative forms used by the interviewer below. The remaining eight are uttered by Diana, Princess of Wales. Again they all feed into the autobiography of the unhappy princess (*…it was the first time I'd experienced what it was like to be outside of the net; …I was the one always pictured…. ; I was the separated wife…, etc.*):

Figure 2 – All of the instances of *was + the* in the Bashir – Diana interview

1	in a serious manner, and that was the first time I'd experi
2	to share that load, because I was the one who was always pi
3	y, yeah. BASHIR: Why? DIANA: I was the separated wife of the
4	from about 1989 I think. What was the nature of your relati
5	confuse the enemy. BASHIR: Who was the enemy? DIANA: Well, t
6	as the cause? DIANA: The cause was the situation where my hu
7	d to be smoothed. BASHIR: What was the family's reaction to
8	ught. The most daunting aspect was the media attention, beca
9	epression? DIANA: Well maybe I was the first person ever to
10	se. DIANA: Uh,uh. BASHIR: What was the cause? DIANA: The cau
11	on a hanger: they decided that was the problem - Diana was u

Was + very

The other main pattern in terms of words that come after *was* is *was + very.* All of these are uttered by Diana and are largely indicative of *very* being used as a modified (in this case an intensifier) of an adjective or a noun phrase. They are again largely relating to the portrayal of the unhappy nature of the life and relationships of the interviewee:

Figure 3 – All of the instances of *was + very* in the Bashir – Diana interview

1	d more cards than I would - it was very much a poker game, c
2	A: No, because again the media was very interested about our
3	I was in love with him. But I was very let down. BASHIR: Ho
4	f fantasy in that book, and it was very distressing for me t
5	mily? DIANA: I think everybody was very anxious because they
6	eople initially? DIANA: Yes, I was very daunted because as f
7	is the bigger the drop. And I was very aware of that. BASHI
8	We struggled a bit with it, it was very difficult; and then
9	t like to fulfil? DIANA: No, I was very confused by which ar

- *What* and questions

We noted above that *what was* accounts for 7% of all of the patterns to the left of the node word *was.* Of the 13 occurrences of the pattern, nine are uttered by the interviewer. Eight of these are wh- interrogatives with *what.*

Figure 4 – All of the instances of *what + was* in the Bashir – Diana interview

1	very busy stopping me. BASHIR: What was your reaction when n
2	orehand, and explained to them what was happening. And they
3	ry good team in public; albeit what was going on in private,
4	you, from about 1989 I think. What was the nature of your r
5	e they were able to understand what was coming out, and I wa
6	ngth from to continue? BASHIR: What was your reaction to you
7	es, a number of times. BASHIR: What was said? DIANA: Well, i
8	e they're coming from. BASHIR: What was your husband's react
9	needed to be smoothed. BASHIR: What was the family's reactio
10	ng before you became pregnant. What was your reaction when y
11	e cause. DIANA: Uh,uh. BASHIR: What was the cause? DIANA: Th
12	he fridge. It was a symptom of what was going on in my marri
13	it. BASHIR: Did he understand what was behind the physical

Wh- interrogatives are prototypical interrogative clauses. They are very much indicative of prepared (and pre-approved) questions as opposed to questions which arise out of the previous response. Compare the follow question and answer sequences in terms of the first comprising of two pre-prepared and planned questions versus the second example where the interviewer's follow up question arise in an ad hoc manner based on the response to the first question:

3) Bashir – Diana, Princess of Wales interview

BASHIR: **What was the family's reaction to your post-natal depression?**

DIANA: Well maybe I was the first person ever to be in this family who ever had a depression or was ever openly tearful. And obviously that was daunting, because if you've never seen it before how do you support it?

BASHIR: **What effect did the depression have on your marriage?**

4) Jeremy Paxman interviewing author JK Rowling on BBC *Newsnight 18 June 2003.*

Full transcript available at:

http://news.bbc.co.uk/1/hi/programmes/newsnight/3004594.stm

JEREMY PAXMAN: **Do you think success has changed you?**

JK ROWLING: Yes.

JEREMY PAXMAN: **In what way?**

JK ROWLING: I don't feel like quite such a waste of space anymore.

JEREMY PAXMAN: **You didn't really feel a waste of space?**

JK ROWLING: I totally felt a waste of space. I was lousy….

The more prototypical the question forms the more formal and pre-scripted they appear. The more ad hoc and less prototypical the question forms, the less pre-scripted they appear. In the Paxman – Rowling interview, there were prototypical wh- questions but there was also a range of other question forms, such as tag questions, declarative questions, double questions. Based on O'Keeffe (2005 and 2006), the comparison of question forms makes for an interesting study of how question forms can vary. In my study I randomly analysed 100 questions in a number of interviews, including Bashir-Diana and Paxman-Rowling as well as data from the TV chatshow Parkinson and a BBC *Newsnight* interview between Jeremy Paxman and the Prime Minister of Britain at the time, Tony Blair. The following range of question types were found:

Table 7 – Range of question types across a 500 questions across media five interviews (O'Keeffe 2005; 2006)

Question type	**Example**
Yes/no	Is it true that you figure it's associated with all sorts of seedy things like venereal diseases or prostitution or that kind of thing?
Wh-	What age is he ah Breda?
Declarative	You won't be seeing the match this weekend?
Double	How did you know? Did the bush telegraph tell you?
Tag	Eh that's the point isn't it?
Alternative	And in terms of changing a climate or an atmosphere ah within the course and within the community within society do you believe it's a legislative requirement or ah a debate requirement?

On analysing the different interviews, the following results were found in relation to the questiontypes across 100 randomly chosen questions in each interview:

Table 8 – Results of analysis of question forms across four TV interviews

Question type	**Bashir– Diana**	**Parkinson (TV chatshow)**	**Paxman– Blair**	**Paxman-JK Rowling**
Yes/no	25	23	30	40
Wh-	40	16	24	18
Declarative	31	39	42	20
Double	3	12	3	15
Tag	0	10	0	7
Alternative	1	0	1	0

O'Keeffe, A., (2012) Corpora and Media Studies. In K. Hyland, M. H. Chau and M. Handford (Eds) Corpus Applications in Applied Linguistics. London: Continuum, pp. 117-131.

From table 8, we can see that the majority of the questions in the Bashir interview fall within the range of the prototypical question forms, yes/no, Wh- and declarative questions. This is also the case in the Paxman-Blair interview which took place 6th February 2003. Full transcript: http://news.bbc.co.uk/1/hi/programmes/newsnight/2732979.stm . In both cases, 97% of all questions asked fall within these three conventional types. This suggests pre-scripting and carefully planning in terms of the interviews. In both cases, these three main question types account for 78% of all the question forms in the Parkinson and Paxman-Rowling interview. In these less formal interviews, there are more ad hoc questions formed which have a more conversational style. In particular we see the use of double questions and tag questions, both of which are much more hedged in nature compared to the more prototypical question forms (which are more face-threatening):

5) Double questions from Paxman-Rowling interview:

Jeremy Paxman: And what about the money? A lot of people when they suddenly make a lot of money, feel guilty about it. Do you feel guilt?

6) Tag question from Parkinson chatshow (interview with Lily Savage)

Michael Parkinson: Well this is your comeback but it's also your final television appearance as Lily Savage, isn't it?

25 December 2004. Full transcripts available at:

http://parkinson.tangozebra.com/guest_transcript.phtml?guest_id=55

O'Keeffe, A., (2012) Corpora and Media Studies. In K. Hyland, M. H. Chau and M. Handford (Eds) Corpus Applications in Applied Linguistics. London: Continuum, pp. 117-131.

Conclusion

This paper has surveyed the core applications of corpus linguistics to the study of media discourse. It has discussed the setting up of a representative corpus and it has exemplified the core corpus functions of key word analysis, word frequency lists and concordancing. These are just the basics of the applications of corpus linguistics to the study of media discourse. As such, the basic application of CL to the study of media texts, whether spoken or written, is to provide handy and quick quantitative analyses that would be difficult and labour intensive, to provide a facility for automated indexed searches of the text. In other words the job of the analysis still remains with the analyst.

This paper has not discussed the wider application of CL in the study of media discourse and that is when it is used in tandem with other analytical approaches and models. There is a growing body of work(s) where CL has been used to mine results within broader analytical frameworks. Generating corpus results on their own describe and quantify the data itself but in order to take these further for the study of media discourse, they need a broader framework. Conversation analysis has provided this framework for most studies of spoken media discourse and Critical Discourse Analysis (CDA) has been the main paradigm for the analysis of written media texts.

In terms of marrying CDA with CL, O'Halloran (2010) provides an excellent example. He shows how CL can be a powerful complementary tool to CDA when he examines a set of texts over a six-week period in the British popular tabloid newspaper *The Sun* on the topic of the European Union (EU) expansion on 1st May 2004. The corpus he built consists of *The Sun* texts in the six weeks prior to 1st May which contain the cultural keywords: '(im)migration', '(im)migrant(s)', 'EU' and 'European'. O'Halloran is able to show, in a convincing and powerful way, how the language and ideology were intertwined in that period. For example, key words such as 'high unemployment', 'impoverished', 'poor' were linked to 'Eastern European' and were tied up with the presupposition that EU enlargement would mean that migrants would be a drain on social services, etc.

In essence, the potential for the use of corpora in the study of language in the media is immense. It allows the analysis of much larger amounts of data in both a

quantitative and qualitative way, as I hope to have illustrated in the brief analyses presented here. However, in order to come to broader conclusions beyond describing how media texts differ from other genres, one needs to work with other discourse frameworks such as CA or CDA.

References

Fairclough, N. (1995a). *Media Discourse,* London: Arnold.
Fairclough, N. (1995b) *Critical Discourse Analysis,* London: Longman
Fairclough, N. (2000) *New Labour, New Language,* London: Routledge & Kegan Paul.
Farr, F., Murphy, B. and O'Keeffe, A. (2004) 'The Limerick Corpus of Irish English: design, description and application', *Teanga,* 21: 5-29.
Hutchby, I. (1991) 'The organisation of talk on talk radio', in P. Scannell (ed) *Broadcast Talk,* London: Sage, pp.119-137.
Hutchby, I. (1996) 'Power in discourse: the case of arguments on a British talk radio show', *Discourse and Society,* 7(4): 481-497.
Malveira Orfanó, B. (2010) *The representation of spoken language: a corpus-based study of sitcom discourse.* Unpublished PhD thesis, Mary Immaculate College, University of Limerick, Ireland.
O'Keeffe (2006) *Investigating Media Discourse.* London: Routledge.
O'Keeffe, A. (2005) ' "You've a daughter yourself?": a corpus-based look at question forms in an Irish radio phone-in', in K.P. Schneider and A. Barron (Eds.) *The Pragmatics of Irish English,* Berlin: Mouton de Gruyter, pp. 339-366.
Scott, M. (2008) *WordSmith Tools version* 5. Liverpool: Lexical Analysis Software.
Sinclair, J. McH. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
O'Halloran, K. (2010) 'How to use corpus linguistics in the study of media discourse', in O'Keeffe, A. and McCarthy, M.J. (eds) *The Routledge Handbook of Corpus Linguistics.* London: Routledge, pp.563-577.