Post-print version of: Gabrielatos, C., McEnery, T., Diggle, P. & Baker, P. (2012). The peaks and troughs of corpus-based contextual analysis. *International Journal of Corpus Linguistics*, 37(2), 151-175. DOI: 10.1075/ijcl.17.2.01gab

The peaks and troughs of corpus-based contextual analysis

Costas Gabrielatos, Tony McEnery, Peter J. Diggle & Paul Baker

This paper focuses upon two issues. Firstly, the question of identifying diachronic trends, and more importantly significant outliers, in corpora which permit an investigation of a feature at many sampling points over time. Secondly, we consider how best to combine more qualitatively oriented approaches to corpus data with the type of trends that can be observed in a corpus using quantitative techniques. The work uses a recently completed ESRC-funded project as a case study, the representation of Islam in the UK press, in order to demonstrate the potential of the approach taken to establishing significant peaks in diachronic frequency development, and the fruitful interface that may be created between qualitative and quantitative techniques.

Keywords: diachrony, brachychrony, sampling points, granularity, quantitative and qualitative techniques

1. Introduction

This study reports on the development of a statistical approach that can aid a 'corpus-based contextual analysis' (see Baker et al. 2008: 295 and Section 5.1 below). As the term implies, the focus of such a study is not on identifying linguistic patterns *per se*, but focuses instead on the identification of relevant contextual elements (e.g. events or actions); therefore, it precedes the linguistic analysis. That is, a contextual analysis seeks to identify salient contextual elements that are relevant to a subsequent linguistic study.[1] In general, the technique seeks to establish statistically significant peaks and troughs (primarily the former) in the number of documents in sub-corpora defined by particular time-points or periods treated as time-points (e.g. by year, month, day of publication).[2,3] Once these time points/periods have been established, the next step is to identify contextual elements which have potentially influenced the increase in the frequency of corpus documents (e.g. events which cause an increase in newspaper articles on a particular topic). For details on the technique used in the case study reported here, see Section 5.1. The contextual elements thus identified have a dual utility:

(a) They point towards sub-corpora that can be usefully examined in more detail using corpus techniques. Alternatively, or additionally, the sub-corpora thus identified can become the source of targeted downsampling, in order to yield a manageable amount of text to be examined using close-reading techniques traditionally used in (critical) discourse analysis (see Wodak & Meyer 2009).
(b) They can aid the interpretation of the linguistic findings − particularly in studies with a historical, social or political focus.

In particular, this paper focuses upon the examination of changes in the number of articles published within a given month in a corpus of newspaper articles on the topic of Islam/Muslims in the years 1998-2009, and the correlation of frequency peaks with events which can be expected to have triggered the increased reporting (see Gabrielatos & Baker 2008: 17-20; Renouf 2002: 35).[4] As such articles were frequent, appearing daily throughout

the period studied, yet also somewhat varied in frequency, the study presented us with an issue: how to track trends in data when the sample points are so frequent and variable. Hence the focus of the study is the discussion of the development of a statistical method for establishing outliers that lie significantly above or below the trend in the data − what we refer to as 'peaks' and 'troughs', respectively. Although the focus of the paper is mainly methodological, the development of a method which establishes peaks and troughs of reporting objectively and with a high degree of confidence has important theoretical implications. The correlations between frequency of reporting and candidate trigger events are taken into account in the interpretation of the results of the corpus analysis – which can then feed into the theoretical frameworks informing critical discourse studies.[5] The utility of the technique also reaches beyond research within academia, as frequency peaks of particular words or topics are increasingly used to establish trends and formulate predictions within political debates. For example, a recent article in the *New York Times* traces changes in the topics seen as salient in the 2012 presidential race with reference to peaks in the frequency of particular (groups of) words used in speeches by Barack Obama and candidates for the Republican nomination between 2009 and 2012.[6] At the same time, the technique can become a useful addition to the methodological toolbox of corpus-based studies with a diachronic focus.[7]

2. Positioning the study: Corpus-based approaches to diachronic analysis

The examination of the diachronic frequency development of lexical or grammatical features is not new within corpus linguistics.[8] For example, Baayen & Renouf (1996) examined the appearance and frequency development of five English derivational affixes *(-ly, -ness, -ity, un-, in-*) in a corpus of the *Times*, spanning four years. Leech (2004) and Leech & Smith (2006) examined the frequency development of modal verbs in British and American English comparing corpora comprising written texts published in the early 1960s and 1990s, whereas Mair et al. (2002) examined the frequency development of part of speech tags in British English over the same period. Leech et al. (2009) expanded the time range adding a corpus of texts published in the late 1920s and early 1930s. Finally, Gries & Hilpert (2008) report on two diachronic studies focusing on the verbal complements of *shall* and the English present perfect, using a number of diachronic corpora spanning over four centuries.

One feature which differentiates diachronic corpus studies is the time-span they cover – which may range from a few years to centuries.[9] However, the time-span may vary simply to suit particular research objectives; therefore, it is not in itself important for the purposes of this paper. A pertinent feature of diachronic studies, we argue, is the number of time points examined within a given time-span. That is, what is crucial for the utility of a diachronic corpus study is the number of time-points at which frequencies have been measured – referred to here as 'sampling points'. The number of sampling points is related to the 'granularity' of the analysis (Davies 2010: 448). The issue of granularity merits attention here, as it is a factor which influences the type of statistical analysis that can be used to identify frequency trends or outliers (see below). As will be shown, the number of sampling points cannot in itself express the granularity of a diachronic corpus study – we also need to take into account the time span of the corpus.

A quick and simple way of quantifying and normalising the granularity of a diachronic study, and thus being able to compare studies in that respect, is to divide the number of sampling points by the time length of the corpus, expressed in a specified reference time unit – say, years.[10] The higher the resulting figure, the higher the granularity of analysis – and the higher

the accuracy and usefulness of the results. For example, Leech & Smith (2009) examine the development of modals in British English within approximately sixty-five years (late 1920s to early 1990s) using three 'snapshot' corpora (Claridge 2008: 243), that is, corpora containing texts published in a particular year or a small number of years. The corpora were, on average, thirty years apart from one another, and were constructed to be representative of English at the time period sampled. Each corpus acted as a sampling point – resulting in a granularity of about 0.05 (3 sampling points divided by the 65 year span). Millar (2009) investigated the frequency development of modal verbs in *TIME Magazine* within a similar time-span (1923–2006, i.e. 84 years) using annual intervals – thus creating 84 sampling points, and a granularity of 1 (i.e. twenty times higher). Even finer granularity was achieved in the study of Baayen & Renouf (1996). Although they examined a significantly shorter period (just under four years), by dividing their corpus into monthly sub-corpora (thus creating 47 sampling points) they achieved a granularity of 12. The latter two studies present a clear example of the inadequacy of measuring granularity by the number of sampling points alone. Although Baayen & Renouf (1996) use about half of the sampling points compared to Millar (2009), their study has twelve times the granularity.

A telling example of the utility of high granularity can be found in Millar (2009). Figure 1 (from Millar 2009: 201) shows the frequency development of the modal *would*. If the development was examined using only three sampling points − say, the years 1923, 1964 and 2006 – then the analysis would show almost no development in the frequency of *would*. As it happens, Millar's statistical analysis returned a similar result (2009: 200-202), as shown by the regression line in Figure 1. However, the frequency development plot (the points in Figure 1) also reveals a clearly pronounced frequency peak in the 1940s – particularly in the second part of the decade − with two more modals (*might, could*) also showing distinct peaks in the same period (Millar 2009: 201). These peaks would remain undetected if the study had used substantially fewer sampling points. As Millar (2009: 200) comments, "the frequency of usage can vary greatly year on year, and these yearly changes often appear to go against the overall trend". Of course, we cannot argue that there is a causal link between World War II and the peak solely on the basis of the correlation between a period of intense uncertainty and a frequency peak. However, neither can a causal link be dismissed out of hand without further investigation. It seems clear, then, that diachronic corpus studies of low granularity run the risk of leaving unrevealed patterns that may be worth further investigation.

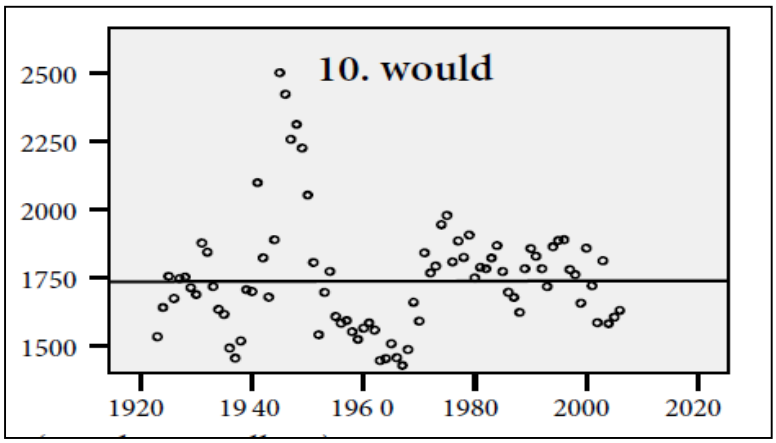

Figure 1. Annual frequency development of *would* in TIME, 1923-2006

A second important aspect of diachronic studies is the use and nature of statistical analyses. That is, it is crucial to establish if diachronic changes (whether overall trends or frequency

differences between two sampling points) were calculated statistically or identified impressionistically, for example, through the examination of graphical representations (see also Gries & Hilpert 2008: 60). In that respect, when statistical analyses are employed, three main approaches seem to be used. The first approach uses cluster analysis to establish diachronic sub-corpora in which the language feature in focus shows comparable frequency (Gries & Hilpert 2008). Although the approach is statistically sophisticated, it is not useful for the aims of the present paper, as its focus is on the detection of frequency similarities, rather than frequency outliers (ibid.: 62-63), that is, its objective is to group rather than isolate time-defined sub-corpora. The second approach treats time points and the frequencies recorded at each point as variables, and carries out regression analyses in order to establish the extent of correlation between the variables of time and frequency (e.g. Baayen & Renouf 1996: 72, 79-81; Millar 2009: 198-202, 217). Simply put, such approaches are concerned with the statistical significance of the overall trend over time. That is, the level of confidence relates to the overall upward or downward frequency development indicated by the regression line fitted on the data. The third approach is concerned with the statistical significance of differences between frequencies recorded at two time points (e.g. Leech & Smith 2006: 188; Leech et al. 2009: 40; Mair et al. 2003: 65-66). It is also pertinent to note that regression analyses or pairwise frequency differences (and their statistical significance) seem to be adopted in diachronic studies of high and low granularity respectively. As will be seen in Section 5, our approach improves on the third approach, while also establishing statistically significant diachronic trends (as in the second approach).

3. The corpus

The corpus used in this study comprises newspaper articles containing one or more of the words in the query below (Table 1) published in 12 national UK newspapers (Table 2) between 1st January 1998 and 31st December 2009.[11,12] The corpus comprises 200,037 articles, amounting to 143 million words. Selecting the query terms posed the following two interrelated challenges (see Gabrielatos 2007):

(a) The query terms should return all articles related, even peripherally or incidentally, to (issues pertaining to) Islam and Muslims, without favouring or filtering out articles expressing particular views towards the religion or its adherents.

(b) The query had to balance recall and precision. More precisely, the query would, ideally, return all relevant articles in the database (i.e. have 100% recall), while avoiding also returning irrelevant articles (i.e. have 100% precision).

Table 1. Query terms[13]

alah OR allah OR ayatolah OR burka! OR burqa! OR chador! OR fatwa! OR hejab! OR imam! OR islam! OR koran OR mecca OR medina OR mohammedan! OR moslem! OR muslim! OR mosque OR mufti! OR mujaheddin! OR mujahedin! OR mullah! OR muslim! OR prophet mohammed OR q'uran OR rupoush OR rupush OR sharia OR shari'a OR shia! OR shi-ite! OR shi'ite! OR sunni! OR the prophet OR wahabi OR yashmak! AND NOT islamabad AND NOT shiatsu AND NOT sunnily

Table 2. Corpus newspapers and abbreviations

The Business (bu)
Daily Express + Sunday Express (ex)
The Guardian (gd)
The Independent + The Independent on Sunday (in)
The Daily Mail + Mail on Sunday (ml)
The Daily Mirror + The Sunday Mirror (mr)
The Observer (ob)
The People (pp)
The Sun + The News of the World (su)
The Daily Star + The Daily Star Sunday (st)
The Times + The Sunday Times (tm)
The Daily Telegraph + The Sunday Telegraph (tg)

It must be stressed that the selection of query terms was not made in ignorance of the issues frequently reported in British newspapers, and the controversy surrounding them (e.g. *headscarf*), as they were the primary motivation for the research carried out in this project (for a discussion on subjectivity and objectivity in corpus-based critical discourse studies, see Baker 2012). However, the terms selected for the query were, at least nominally, only descriptive of the religion, its believers, and attendant practices (e.g. related to worship). The only exceptions were the terms *Mohammedan(s)* and *Moslem(s)*, both derogatory terms referring to Muslims, which were included so that it would be possible to examine their frequency and distribution if needed. Their inclusion is not expected to have influenced the results of the analysis, as their frequency is very low: *Mohammedan(s)* is only found 77 times in the corpus, and *Moslem(s)* 5,832. The low frequency of the latter is better understood in relation to the frequency of *Muslim(s)*: 127,598.[14] Even incidental mentions of the query terms were deemed relevant as their use in any context contributes to, and reveals, the general picture of the presentation of Islam and Muslims in the British press (Gabrielatos 2007: 10-11).

The corpus was divided into sub-corpora in terms of time and newspaper (i.e. a sub-corpus per month per newspaper) to facilitate the investigation of diachronic development of patterns, or patterns specific to particular periods, both in the whole corpus and particular newspapers. This allows the examination of patterns in periods of markedly increased publication of relevant articles (see below for more details). Let us now outline the characteristics of the present study.

4. Characteristics of the study: Theoretical and methodological underpinnings

As the query is topic-related, a significantly increased frequency of articles is expected to correspond to increased reporting of the topic. The higher volume of reporting can be seen to indicate either rising public interest in the topic, or editorial/management decisions to create high visibility for it. In either case, the frequency of reporting of an event or topic can be safely treated as proportionate to its actual, perceived or projected salience. This is particularly important for the approach informing the current study, as well as the one in which the approach discussed here was piloted (see Baker et al. 2008; Gabrielatos & Baker 2008), namely Critical Discourse Analysis (CDA), and in particular the Discourse-Historical approach within CDA (see Reisigl & Wodak 2001: 31-90). In this approach, the analysis takes into account the following elements (ibid.: 40-41):

(a) The immediate co-text (e.g. collocations and resulting semantic prosodies).
(b) Intertextual and interdiscursive relations.
(c) Relevant contextual elements; e.g., "the occasion of the communicative event" (ibid: 41).
(d) The broader sociopolitical and historical context.

The approach discussed here aims at objectively identifying element (c), in that the establishment of significant peaks of reporting can lead to the identification of candidate trigger events, which, in turn, can illuminate the relevant contextual background. The latter can then greatly assist the interpretation of corpus findings (e.g. through collocational networks), as well as objectively pinpoint time periods within which texts can be selected for qualitative critical discourse analysis – mainly through data downsampling (see Baker et al. 2008: 284-285; KhosraviNik 2009, 2010).

The time span of 12 years (1998 – 2009) is particularly short, and lies at the lower end of the span of 10–30 years that Mair (1997) termed 'brachychrony'.[15] However, our study is among those with the highest granularity (see Section 2 above), as dividing the corpus into monthly intervals (thus creating 144 sampling points) achieved a granularity of 12. More importantly, our study differs from those usually carried out within corpus linguistics (see above for examples) in three respects:

(a) Usually, corpus-based diachronic studies are carried out on corpora of samples (i.e. texts selected to provide a representative picture of a language or genre). However, our corpus represents the entirety of newspaper articles mentioning the query terms published in the British national press over the specified time period.
(b) Corpus-based diachronic studies focus on the development of particular lexical or grammatical constructions – whereas we are examining the frequency development of articles containing one or more of the query terms (all having to do with a specific religion (Islam) and its adherents).
(c) The focus is not so much the establishment of diachronic trends, as the identification of extreme points in the frequency development – namely, distinct frequency peaks and troughs (see above).

This approach was initiated in a previous project examining the presentation of refugees, asylum seekers, immigrants and migrants (RASIM) in the UK press, in order to identify peaks in the reporting (for details, see Baker et al. 2008).[16] The analysis had very high granularity: 132 monthly sampling points over eleven years, resulting in a granularity of 12. However, it was carried out in a limited and statistically naïve manner (see Gabrielatos & Baker 2008: 17-20) as:

(a) the frequency development of query terms was examined only in the whole corpus (i.e. all twelve corpus newspapers were taken collectively),
(b) what was plotted was the average number of articles per month, and
(c) the peaks were established impressionistically (i.e. by examining the plot), and, therefore, only what appeared to be very pronounced peaks were taken into account.

Nevertheless, the identification of major peaks and the corresponding trigger events was indeed useful to the CDA strand of the above project in two respects. It enabled the application of "a preliminary restrictive factor in downsampling the texts" (KhosraviNik 2010: 5). It also rendered the downsampled data to be analysed qualitatively "sensitive to the

aims of deconstructing the representation of RASIM in the context of relevant socio-political developments, instead of applying a randomized text selection" (KhosraviNik 2009: 482).

A similar approach was initially adopted in the project reported on here; however, the frequency development of articles over time was examined not only on all corpus newspapers collectively, but also on each newspaper individually. A critical summary of the initial analysis is presented in Section 5.1. This outlines the process of 'contextual analysis' – that is, the linking of peaks in reporting to particular world events. It also highlights the merits of the methodology reported here; for example, the significant peaks were predominantly those shared by the vast majority of newspapers in the corpus. The problems which prompted the development of the methodology are reported in Section 5.2.

5. Analysis and discussion

5.1 Initial corpus-based contextual analysis

The first stage of the analysis examined the diachronic development of the total number of corpus articles for each newspaper in each month in order to:

(a) identify time points (in our case, periods of one calendar month) showing significant increases (peaks) in reporting;
(b) connect these peaks to potential *trigger events*, that is events which are likely to have had increased coverage (in terms of number of articles) in the corpus newspapers (Table 3 – bold indicates peaks shared by all/most corpus newspapers). These events were identified by (i) reading a sample of articles from the particular time period and newspaper which constituted a peak, and (ii) entering the query (see Table 1 above) in Google News, using the "custom range" function to limit results to those published in each month showing a peak in our corpus, and examining the results for news stories repeated frequently (the procedure was repeated for each month showing a spike − whether established impressionistically or through the WST method).[17]

At this point we need to address the issue of identifying events connected to troughs. While the process connecting peaks to candidate trigger events is perhaps fairly impressionistic, but feasible, it is much less straightforward in the case of troughs. This is because significant troughs may be due to the absence of relevant trigger events (perhaps in combination with media 'topic fatigue' following a period of intense reporting on a particular topic), or important events unrelated to the topic, which caused the focus of reporting to shift from the topic investigated in the study. Linking troughs to either absence of reporting or increased reporting on another topic can only be speculative, and quite difficult to support. Clearly, techniques need to be developed which address the issue in a satisfactory way; however, in their current absence, the paper will only report on the link between peaks and candidate trigger events.

Table 3. Peaks, as identified impressionistically − and the corresponding candidate trigger events

#	Peak date	Trigger Event	*bu*	*ex*	*gd*	*in*	*ml*	*mr*	*ob*	*pp*	*st*	*su*	*tg*	*tm*
1	1999, Jan.	• Iraq bombing in December 1998 (Operation Desert Fox).[18] • Christian-Muslim clashes in Indonesia.	✓											
2	2000, June-July	• Christian-Muslim clashes in Indonesia. • Muslim rebels in the Philippines take hostages.	✓											
3	2000, Sept.-Oct.	• Muslim rebels in the Philippines take hostages.	✓											
4	**2001, Sept-Oct.**	• 9/11.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
5	2002, March-April	• Hindus attacking Muslims in Western India. • Discussion on the status of Muslims in the US. • Reports on female genital mutilation in Muslim countries. • Stoning of woman in Nigeria.							✓					
6	2003, March-April	• Invasion of Iraq.			✓	✓		✓		✓				✓
7	2004, March-April	• First anniversary of Iraq invasion. • Madrid bombing.		✓						✓		✓		
8	**2005, July-Aug.**	• 7/7.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
9	**2005, October**	• Bali bombings. • Publication of Prophet Mohammed cartoons (C).	✓	✓	✓	✓	✓	✓	✓	✓	✓			
10	2006, Jan-Feb.	• Protests in Muslim countries in reaction to the publication of the Prophet Mohammed cartoons.					✓		✓			✓		
11	**2006, Oct. (-Dec)**	• Straw comments on Muslim women wearing the veil(Veil).[19] • Prophet Mohammed cartoons – first anniversary (C2).	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓
12	2007, Jan.-Feb.	• Somalia: Islamic rebels (Union of Islamic Courts) vs. government forces.									✓	✓		
13	2007, July	• No particular relevant event stands out.					✓					✓		
14	2007, Oct.-Dec.	• No particular relevant event stands out.								✓	✓			
15	2008, Feb. - March	• Archbishop's comments on Sharia law.[20]							✓			✓		
16	2008, July	• No particular event stands out in frequency.										✓		

#	Peak date	Trigger Event	*bu*	*ex*	*gd*	*in*	*ml*	*mr*	*ob*	*pp*	*st*	*su*	*tg*	*tm*
17	2009, March-April	• Obama visits Turkey.			✓							✓		
18	2009, June	• Iran elections.			✓							✓		✓
19	2009, August	• Flogging penalty for alcohol drinking in Malaysia. • Clashes in Thailand. • Clashes in Somalia.							✓					

The examination of the frequency development of query terms in individual newspapers showed that the newspapers in the corpus do share some trends. First, all corpus newspapers but one show an upward trend. Second, four peaks are shared by at least two-thirds of the newspapers (parentheses indicate the number of newspapers demonstrating the peak): 9/11 (12), 7/7 (12), Veil and/or Cartoons (11), Bali bombings and/or Cartoons (9). As a result, all four peaks are prominent in the frequency development in the whole corpus (Figure 2).

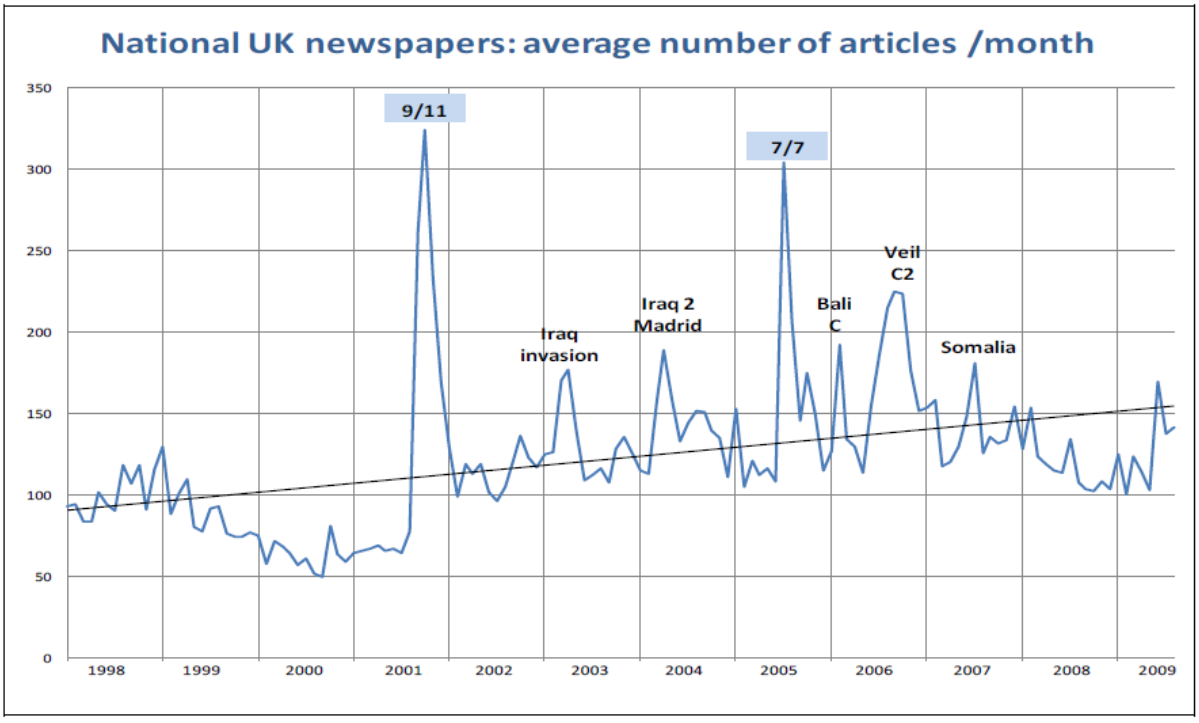

Figure 2. Frequency development of query terms in the corpus, with main potential trigger events indicated[21]

However, the corpus newspapers differ from one another in four interrelated respects. First, although 19 peaks have been collectively identified, only 5 are shared by more than half of the newspapers. Second, different newspapers show different numbers of peaks. Third, the development in the number of corpus articles differs from newspaper to newspaper – something clearly indicated by the different shapes of the lines depicting the frequency development in each corpus newspaper (compare for example, Figures 4, 5 and 7 which show graphs of the numbers of articles per month for *The Guardian*, *The Mirror*, *The Sun* and *The Observer*). This observation also questions the utility of examining the development in the number of articles in the corpus as a whole – thus effectively treating British national newspapers as a homogeneous group. Finally, even when newspapers share peaks, they differ in the degree of their response to the same trigger events – in terms of the number of relevant

articles published in the month containing the event, or the month following it if the event took place at the end of a month or spanned two months (see Gabrielatos 2009). When only major peaks are considered, the corpus newspapers can be divided into five groups according to the most pronounced peaks they feature.[22]

- *9/11 and 7/7*. It is interesting that, although both of these events corresponded to very clear peaks in the frequency development of query terms in the whole corpus (see Figure 2), they were prominent together in only three individual newspapers: *The Guardian, The Independent* and *The Mirror*. It is also noteworthy that these newspapers differ in their response to other candidate trigger events. For example, out of the three, only *The Independent* shows a clear peak corresponding to the veil controversy.
- *9/11 only*. This group is the most populous, comprising four corpus newspapers: *The Business, The People, The Telegraph, The Times*.
- *7/7 and the veil*. Only two newspapers demonstrate this pattern: *The Express* and *The Observer*.
- *Other peaks (9/11 and 7/7 peaks are less prominent)*. This group comprises *The Daily Mail* and *The Sun*.
- *Other peaks (no peaks for 9/11 and 7/7)*. Only *The Star* shows this pattern.

The above comparison of the diachronic development of query terms makes it evident that prominent peaks in the corpus as a whole may not always correspond to peaks present at that point in all individual newspapers. This is because the proportions of articles from each newspaper are unequal in the corpus. As Figure 3 shows, three out of the twelve newspapers (*The Times, The Independent* and *The Guardian*) account for half the corpus articles, while half the corpus newspapers (*The Sun, The Express, The Observer, The Star, The People* and *Business*) account for under a quarter of the articles (22%). Therefore, it is not unreasonable to expect that the trends of the former group of newspapers have influenced those exhibited by the corpus as a whole (Figure 2). Furthermore, even within the above groups, newspapers differ in their rate of frequency development of articles and/or the number and identity of lower-order peaks. Therefore, their overlap in prominent peaks does not seem to be a reliable indicator of similarities in other diachronic trends. Furthermore, none of the groups above is homogeneous in terms of the broadsheet-tabloid distinction, hence the distinction is not explanatory of the results. In light of the above, studies of groups of newspapers, taken as a whole, may miss important individual differences. Conversely, studies of individual newspapers can safely generalise only about the particular newspaper.[23] Therefore, if the corpus comprises distinct sub-corpora (in our case, different newspapers), then frequency developments should be examined in those individually as well as in the corpus as a whole.

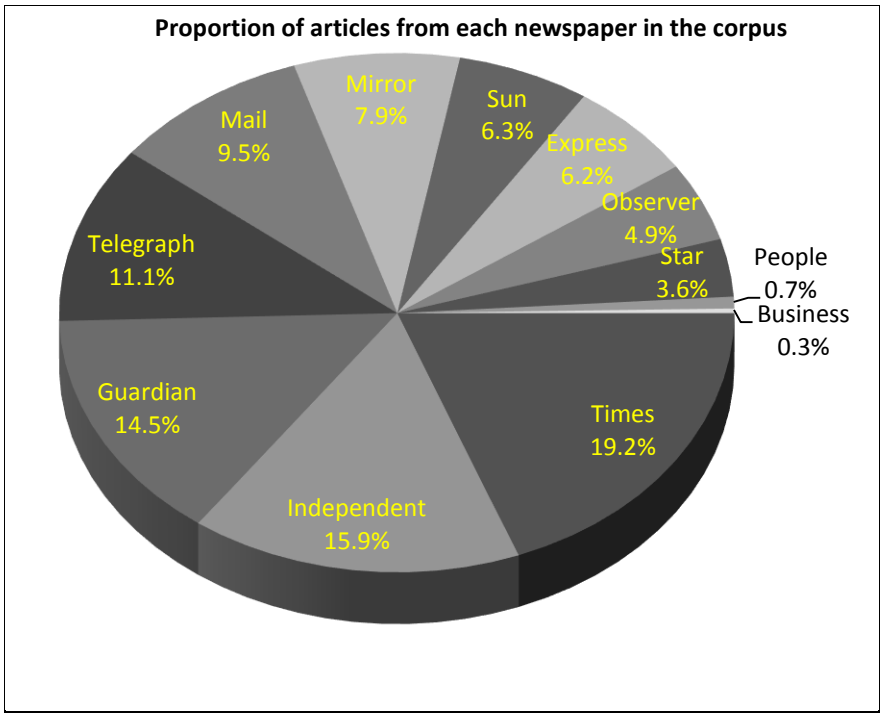

Figure 3. The proportion of corpus representation of each newspaper

The utility of the contextual analysis described above lies in its ability to reveal helpful contextual information that cannot be safely established through introspection. Such information can guide and aid the subsequent linguistic analysis in two respects: (a) it can pinpoint time periods, sources (e.g. newspapers), or texts that can be usefully examined in detail, (b) by revealing elements of the relevant contextual background, it can guide the interpretation of findings (for examples, see Baker et al., forthcoming (a), (b)). However, the procedure described in this section, helpful as it might be, is fairly unsophisticated, as it does not account for a number of variables – most significantly, the number of articles at the sampling point prior to the candidate peak. The following section discusses the statistical approach adopted in the identification of statistically significant peaks. The results of the statistical analysis clearly demonstrate the shortcomings of only examining the development in the actual or normalised frequency of articles, and manually identifying prominent peaks.

5.2 A statistical approach to establishing frequency peaks/troughs in diachronic corpus studies

The most important observation for the purposes of this paper is that not all peaks are equal – for example, a peak right after a trough shows a more pronounced frequency change than a peak following an average frequency point, and more so than one following a low-level peak. However, despite the variability revealed, the analysis seems to strongly support the approach of linking peaks to events, as in the vast majority of cases the time points of peaks do correspond to periods around significant events.

The analysis of the development in the raw frequency of articles does not take into account two interrelated features. First, in almost all of the corpus newspapers, the number of articles related to Islam and/or Muslims has been increasing since 9/11. Second, there are clear fluctuations in the output of relevant articles between and within newspapers. As a result of the above, peaks that seem to be equally high (i.e. their highest points represent comparable frequencies) may not be of the same salience, because they may start from higher/lower levels

of output. For example, both *The Guardian* and *The Mirror* share the 7/7 peak, with about 450 articles in both newspapers (Figure 4). However, these peaks cannot be deemed equally important. In *The Guardian*, the previous sampling point has a frequency of 250, whereas in *The Mirror* it has a frequency of only 75. That is, the proportions of frequency increase between the two sampling points are 80% and 500% respectively. Clearly, the 7/7 peak is much more dramatic in *The Mirror* than *The Guardian*.

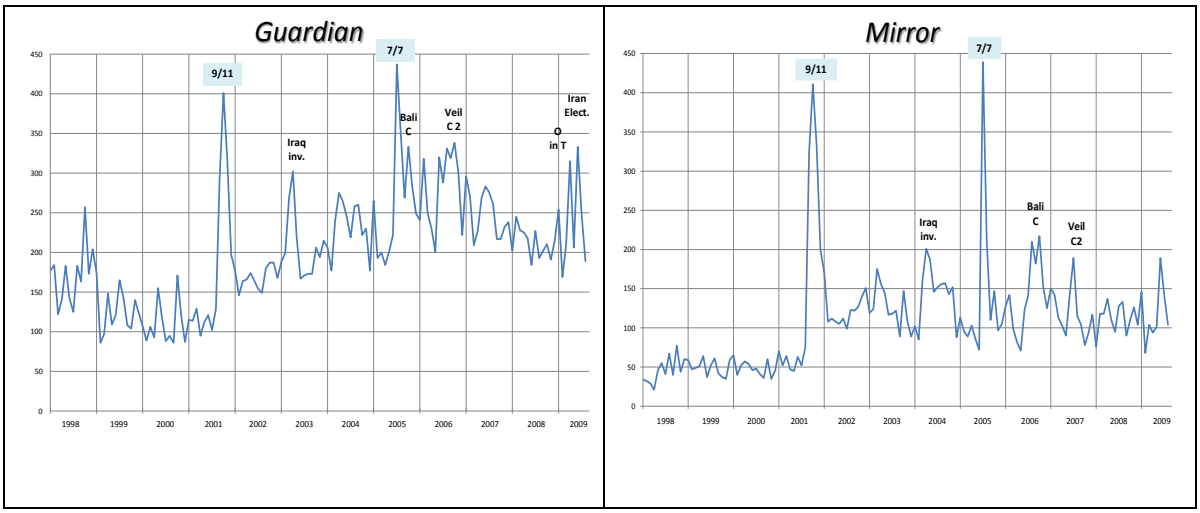

Figure 4. Comparison of the 7/7 peak in *The Guardian* and *The Mirror*.

In order to factor in the issues discussed in Section 5.1 above, and to examine the development of the relevant output of all newspapers in a way that facilitates the comparison of results in a statistically rigorous manner, we analysed the data as follows.

For each newspaper in each month, we converted the raw frequency to a relative change by calculating the logarithm of the frequency difference with the previous month and plotted this variable, $Y(t)$, against time in months, t. In this way, a frequency peak of a given size is considered to be more salient if it follows a trough than if it forms part of a sustained sequence of high frequencies. To distinguish between statistically significant and non-significant peaks, we then fitted the following non-parametric regression model:

$$Y(t) = s(t) + Z(t)$$

where $s(t)$ is a smooth function of time, to be estimated, and $Z(t)$ is a sequence of independent error terms. The model was fitted using the *mgcv* package (Wood 2006) within the R software environment.[24] The output from the package includes estimates and standard errors for the function $s(t)$, from which we construct 95% and 99% point-wise confidence limits as the estimate plus and minus 1.96 and 2.58 standard errors, respectively. Salient peaks were then identified as those that lay outside the 99% confidence limits; however, peaks that lay inside these limits, yet outside the 95% limits, were also considered, particularly when present in a large number of newspapers. In the following graphs, significant peaks are denoted by stars and triangles, corresponding to their lying outside the 99% or 95% confidence limits, respectively. For ease of reference, we will call this technique the wave, peak and trough (WST) method, as it identifies general trends as well as statistically significant peaks and troughs.

The statistical analysis of month-by-month normalised changes in the number of articles revealed that statistically significant peaks were a markedly smaller subset of those identified

in the manual analysis based on the raw frequency of articles. Table 4 provides a summary comparison of the different (number of) peaks identified through the manual and WST analyses – a ✓ indicates peaks established impressionistically, whereas shaded cells indicate peaks established statistically. Overall, there were no more than four WST-identified peaks in each newspaper, with two-thirds of the newspapers (8) having no more than two. Two events emerged as being by far the most salient regarding the volume on reporting: (a) 9/11 – a significant peak in all 12 newspapers, and (b) 7/7 – a significant peak in 9 newspapers (three-quarters of all newspapers). Only five other events showed peaks (cartoons, veil, Iran elections, Iraq invasion, Muslim rebels in the Philippines) – and in only five newspapers, collectively. In other words, only two peaks can be deemed as characterising the national UK press as a whole, the rest being newspaper-specific. At the same time, the analysis confirmed that each newspaper demonstrates a unique reaction to trigger events – depicted in the distinctly different shapes of the WST curves. Also, the analysis revealed a number of significant troughs in all but two newspapers (*Mirror*, *Independent*). It is also interesting that one peak (September-October 2000) which was manually identified in one newspaper (*Business*) proved to be statistically significant in another (*Telegraph*), although it was not manually identified in the latter. Finally, a peak manually unidentified (November –December 1998) was shown to be statistically significant in one newspaper (*Guardian*).

Table 4. Comparison of peaks identified manually and through the WST approach

#	Peak date	*bu*	*ex*	*gd*	*in*	*ml*	*mr*	*ob*	*pp*	*st*	*su*	*tg*	*tm*
1	1998, Nov.-Dec.												
2	1999, January	✓											
3	2000, June-July	✓											
4	2000, Sept.-Oct.	✓											
5	2001, Sept-Oct.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
6	2002, March-April							✓					
7	2003, March-April			✓	✓		✓		✓				✓
8	2004, March-April		✓						✓		✓		
9	2005, July-Aug.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
10	2005, October	✓	✓	✓	✓	✓	✓	✓	✓	✓			
11	2006, Jan-Feb.					✓		✓			✓		
12	2006,Oct.(-Dec)	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓
13	2007, Jan.-Feb.									✓	✓		
14	2007, July					✓					✓		
15	2007, Oct.-Dec.								✓	✓			
16	2008, Feb. - March							✓			✓		
17	2008, July										✓		
18	2009, March-April			✓							✓		
19	2009, June			✓							✓		✓
20	2009, August							✓					

As an example, let us compare the results of the two analyses using *The Sun* – as it was the newspaper which registered the most peaks in the manual analysis (Figures 5 and 6). The first observation is that the two approaches indicated different number of peaks. The raw-frequency approach indicated three major peaks (Veil, 7/7, 9/11) and eight minor ones. In contrast, the WST approach revealed only two statistically significant peaks (9/11, 7/7). Although there is overlap between the major peaks in the raw-frequency approach and the WST peaks, there are important differences. To start with, *the Veil* is the most pronounced peak in Figure 5 (established by its frequency), whereas it is absent in Figure 6. The most salient peak in Figure 6 is 9/11. In fact, not only does September 2001 show a peak, but so do October and November 2001 – the former two being highly significant. In contrast, 7/7 shows only one significant peak ($p<0.05$).[25]

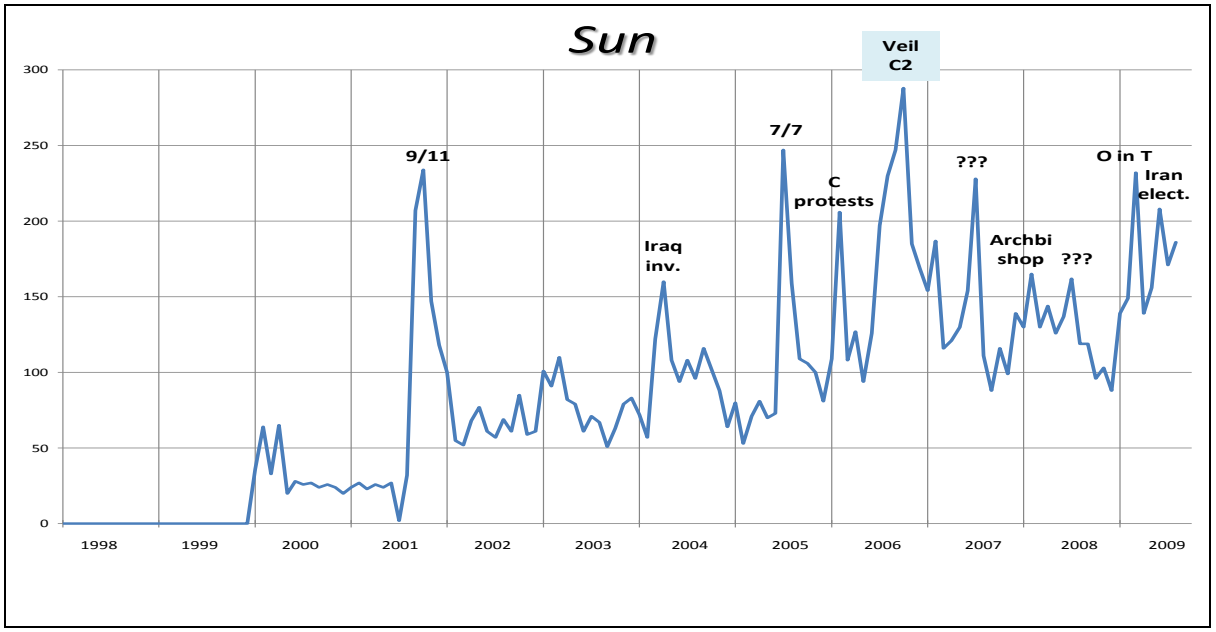

Figure 5. Development of raw frequencies of articles in *The Sun*

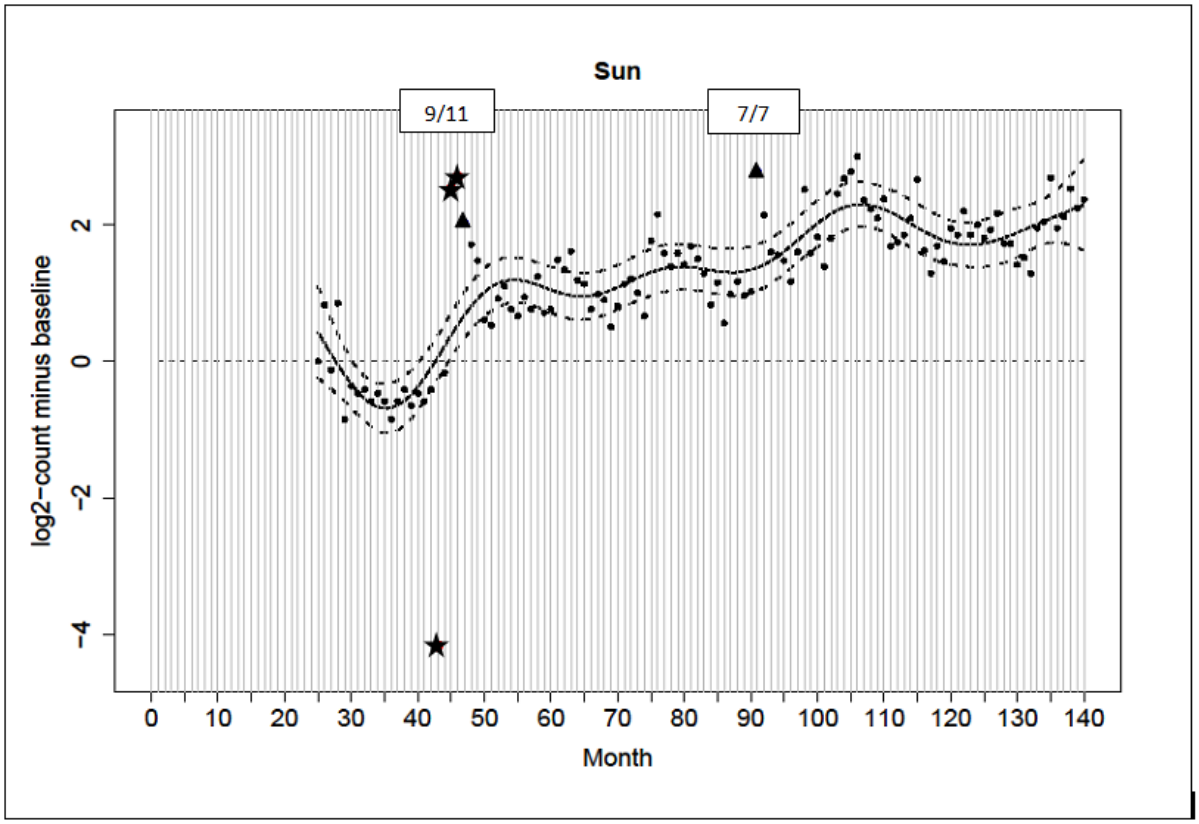

Figure 6. Development of the WST of month-by-month differences in the frequency of articles in *The Sun*

The statistical approach discussed in this paper seems to drastically reduce the number of peaks that may be established impressionistically. This has multiple and interconnected benefits for corpus-based critical discourse studies. First, it limits the periods within which downsampling needs to be carried out for close analysis − either of the type traditionally used in CDA, or the detailed analysis of (expanded) concordance lines carried out within corpus-

based approaches to critical discourse studies. This not only helps focus the analysis (see also below), but also helps save on research time and costs. Second, the interpretation of both corpus and CDA findings draws only on the salient contextual elements (e.g. trigger events, their background, and other connected events), that is, non-salient elements can be excluded from the interpretation. Seen from a different angle, non-salient contextual elements, which might be deemed relevant introspectively, will not be allowed to skew the conclusions. Finally, contextual elements can be weighted in three ways, by treating the following as complementary measures of the salience of a particular event:

(a) The *p*-value of a given peak. For example, peaks at the 99% level of confidence can be treated as more salient or dependable than those at the 95% level.
(b) The proportion of newspapers in which a particular event has registered a peak. For example, in our study, 9/11 registered a peak in all newspapers.
(c) The number of peaks that a particular event registers, either overall in the corpus, or in particular newspapers. Events registering multiple peaks can be treated as more salient than those registering only one. A clear example from our study is 9/11, which registered multiple peaks in a number of newspapers.

A further useful insight for diachronic corpus-based critical discourse studies is that once an influential event has resulted in a dramatic increase in the number of articles reporting on (an aspect of) its topic, and has, additionally, established a consistently higher post-event (and post-peak) output of relevant newspaper articles (as in the case of 9/11), it may take an event of the same, if not greater, magnitude for a statistically significant peak to be established later. This is because the post-peak higher level of topic-related reporting requires that any subsequent peaks represent a significant increase in the number of articles. Put simply, such events raise the bar for subsequent candidate peaks, unless, for example, troughs in due course reduce the height of that notional bar. In addition, the comparisons of the raw counts and WST graphs in our study suggest that, in such cases, the post-event article output need not be smooth, but can demonstrate considerable fluctuations. Let us take the example of the *Observer*, the fluctuating output of which resembles a succession of peaks and troughs. The manual examination of the development of raw counts established eight peaks (Figure 7). Among them, 7/7 and, to a lower degree, the protests on the cartoons of the Prophet Mohammed, and Jack Straw's comments on the veil established the three major peaks – with 9/11 only showing a middle-ranking peak. However, only three of the eight peaks proved to be statistically significant, with 9/11 being one of them (Figure 8). As in almost all newspapers in our corpus, 9/11 signalled an increase in reporting on Islam and Muslims.[26] As a result, only two of the seven subsequent manually-established peaks were statistically significant, and only one (7/7) at the same level of significance as 9/11 ($p<0.01$), the other (cartoon protests) being significant at $p<0.05$. What is pertinent to observe is that three of the non-significant peaks have raw outputs comparable to, or slightly higher (up to 13%) than that of 9/11. The only post-9/11 peak with the same significance as 9/11, that is, 7/7, shows a raw output 37% higher than 9/11.

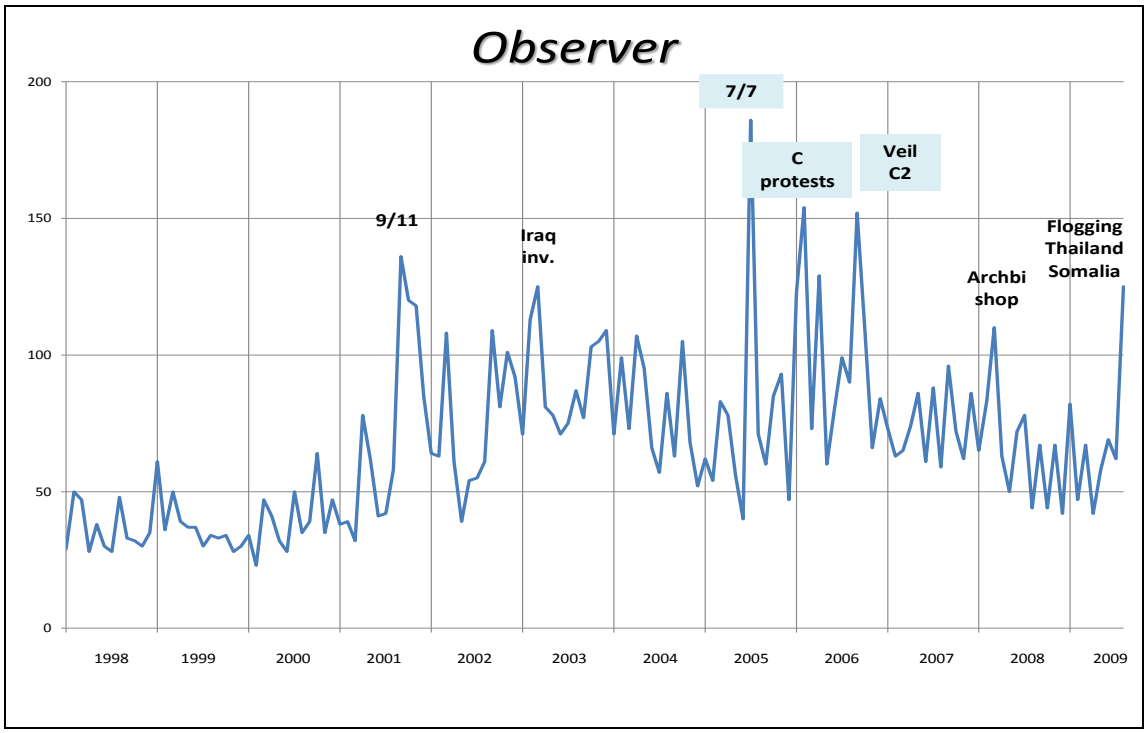

Figure 7. Development of raw frequencies of articles in *The Observer*

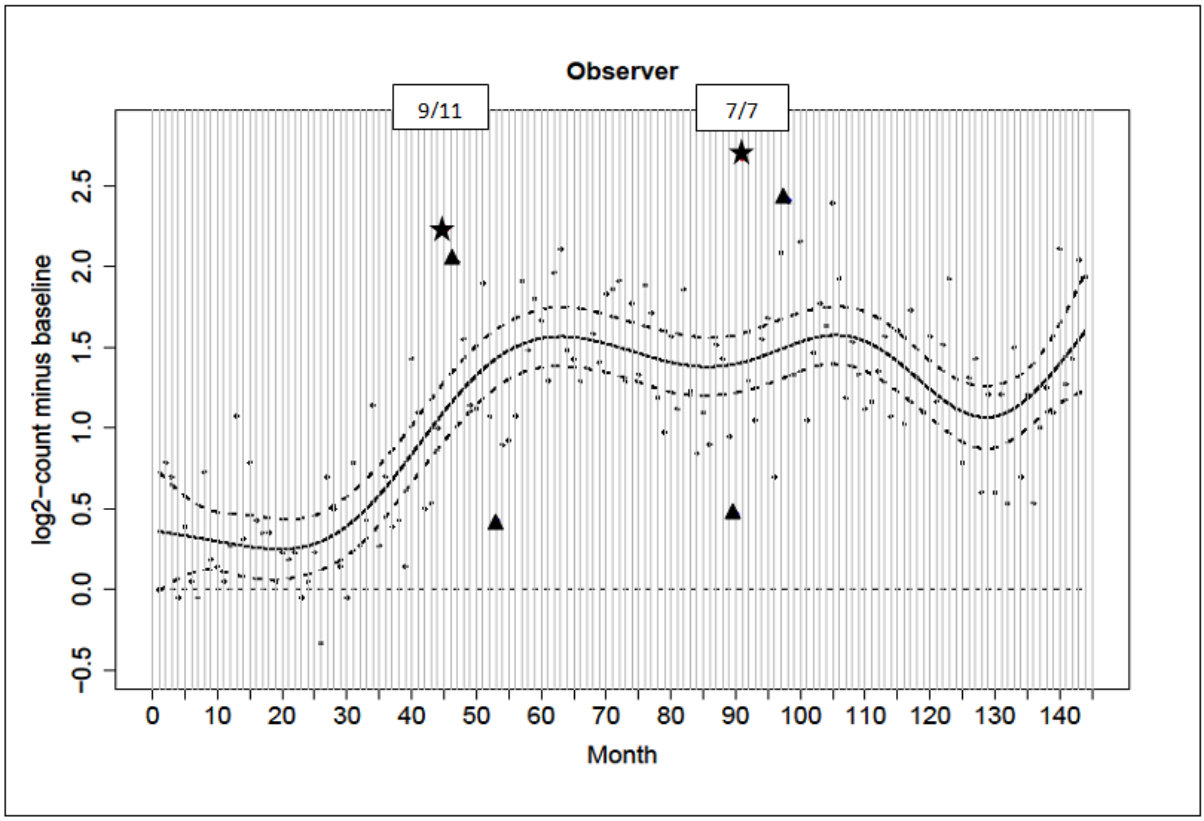

Figure 8. Development of the WST of month-by-month differences in the frequency of articles in *The Observer*

The analysis here has focussed on changing numbers of articles over time. However, there is no reason why the same technique could not be used to analyse change over time in linguistic features such as the frequency of a particular word or multi-word unit, or a set of words

constituting a grammatical or semantic category. Additionally, the WST method can be applied to studies of any level of granularity – although studies of high granularity stand to benefit more, as the technique helps establish extreme fluctuations more accurately than a manual analysis. Irrespective of the focus of the study and the methodology used in the data analysis, the technique outlined here can be usefully applied to the initial examination of the data. Not only does it provide the analyst with a map of the frequency development of the feature in focus, but it also accurately pinpoints extreme values (peaks and troughs). Therefore, it may be used not only in the initial stages of corpus-based critical discourse studies, but also form an integral part of studies focussing on linguistic features. A case in point is the peak established in the use of *would* in Millar (2009), as was shown in Section 2 (see also Figure 1). As the peak was registered towards, and immediately after, the end of World War II, possible research questions arising from the observation could be whether this event motivated increased use of particular modal notions among those expressed by *would*, and whether this putative use was linked to the performance of particular functions (e.g. hypothesising about past events).

6. Conclusion

The increasing availability of texts in electronic form allows for corpora to be constructed which contain many more sampling points over time than has been possible before. The ability to access newswire archives and newswires in real time allows for the analysis of long stretches of time with a granularity that can easily reach the level of daily sampling if required. Web-based corpora also allow increasingly for diachronic studies of unprecedented scope and granularity. Such changes need new techniques to be developed to deal with the complexity that such data expose the analyst to – this paper proposes one such technique. However, the technique reported also responds to another pressure – the need to downsample from large corpora in order to undertake focussed qualitative analyses. This pressure comes especially from linguists wanting to adopt a corpus-based approach, but who wish to combine that with a more nuanced study of a smaller number of texts. In this paper we have identified the need to do this in corpus-based discourse studies, but the same pressure could easily be exerted by other analysts using corpora, notably sociolinguists. Again, this paper proposes a technique which is of use to such users. A salient feature of the development of corpus linguistics has been the development and adoption of ever more sophisticated techniques for manipulating quantitative data derived from corpora. As the nature of corpora have changed, and as the corpus has been adopted to explore an ever wider set of research questions, so the techniques used by corpus linguists to analyse the corpus have been forced to develop. This paper is a contribution in that tradition.

References

Baayen, R.H. & Renouf, A. (1996). Chronicling the times: Productive lexical innovations in an English newspaper. *Language*, 72(1), 69-76.

Baker, P. (2012). Acceptable bias? Using corpus linguistics methods with critical discourse analysis. *Critical Discourse Studies*, 9(3), 247-256.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, P., Gabrielatos, C. & McEnery, T. Forthcoming (a). *Discourse Analysis and Media Attitudes: The representation of Islam in the British press*. Cambridge: Cambridge University Press.

Baker, P., Gabrielatos, C. & McEnery, T. Forthcoming (b). Sketching Muslims: A corpus-driven analysis of representation around the word "Muslim" in the British press, 1998-2009. *Applied Linguistics*.

Baker, P., Gabrielatos C., KhosraviNik, M., Krzyzanowski, M., McEnery, T. & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society,* 19(3), 273-305.

Claridge, C. (2008). Historical corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An international handbook, Volume 1*. Berlin/New York: Mouton de Gruyter, 242-259.

Gabrielatos, C. (2009). Corpus-based methodology and critical discourse studies: Context, content, computation. *Siena English Language and Linguistics Seminars* (SELLS), University of Siena, 9 November 2009. [Abstract and slides available online: http://eprints.lancs.ac.uk/28460/]

Gabrielatos, C. (2007). Selecting query terms to build a specialised corpus from a restricted-access database. *ICAME Journal*, 31, 5-43. [Also online: http://icame.uib.no/ij31/ij31-page5-44.pdf]

Gabrielatos, C. & Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996-2005. *Journal of English Linguistics*, 36(1), 5-38.

Gries, S.Th. & Hilpert, M. (2008). The identification of stages in diachronic data: Variability-based neighbour clustering. *Corpora*, 3(1), 59-81.

KhosraviNik, M. (2009). The representation of refugees, asylum seekers and immigrants in British newspapers during the Balkan conflict (1999) and the British general election (2005). *Discourse and Society*, 20(4), 477-498.

KhosraviNik, M. (2010). The representation of refugees, asylum seekers and immigrants in the British newspapers: A critical discourse analysis. *Journal of Language and Politics*, 8(3), 1-29.

Kytö, M. (1996). *Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding conventions and lists of source texts* (3rd ed.) Helsinki: Helsinki University Printing House. [Also online: http://khnt.hit.uib.no/icame/manuals/hc/index.htm]

Leech, G. (2004). Recent grammatical change in English: Data, description, theory. In K. Aijmer & B. Altenberg (Eds.), *Advances in Corpus Linguistics: Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23) Göteborg 22-26 May 2002*. Amsterdam: Rodopi, 61-81. [Also online: http://www.lancaster.ac.uk/fass/doc_library/linguistics/leechg/leech_2004.pdf]

Leech, G. & Smith, N. (2006). Recent grammatical change in written English 1961-1992: Some preliminary findings of a comparison of American with British English. In A. Renouf & A. Kehoe (Eds.), *The Changing Face of Corpus Linguistics*. Amsterdam: Rodopi, 186-204. [Also online: http://www.lancaster.ac.uk/fass/doc_library/linguistics/leechg/leech_and_smith_2006.pdf]

Leech, G., Hundt, M., Mair, C. & Smith, N. (2009). *Change in Contemporary English: A grammatical study*. Cambridge: Cambridge University Press.

Mair, C. (1997). Parallel corpora: A real-time approach to the study of language change in progress. In M. Ljung (Ed.) *Corpus-based studies in English*. Amsterdam/Atlanta: Rodopi, 195-209.

Mair, C., Hundt, M., Leech, G. & Smith, N. (2003). Short term diachronic shifts in part-of-speech frequencies: A comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics*, 7(2), 245-264.

Mair, C. & Leech, G. (2006). Current change in English syntax. In B. Aarts & A. MacMahon (Eds.) *The Handbook of English Linguistics*. Oxford: Blackwell, 318-342.

McEnery, T. & Gabrielatos, C. (2006). English corpus linguistics. In B. Aarts & A. McMahon (Eds.) *The Handbook of English Linguistics*. Oxford: Blackwell, 33-71.
McEnery, A.M., Xiao, R.Z. & Tono, Y. (2006). *Corpus-based Language Studies : An advanced resource book.* London: Routledge.
McEnery, T. & Hardie, A. (2012). *Corpus Linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.
Millar, N. (2009). Modal verbs in TIME: frequency changes 1923-2006. *International Journal of Corpus Linguistics,* 14(2), 191-220.
Partington, A. (Ed.) (2010). *Modern Diachronic Corpus-assisted Discourse Studies*. Special Issue of *Corpora*, 5(2). Edinburgh: Edinburgh University Press.
Reisigl, M. & Wodak, R. (2001). *Discourse and Discrimination: Rhetorics of racism and antisemitism*. London: Routledge.
Renouf, A. (2002). The time dimension in modern corpus linguistics. In B. Kettemann & G. Marko (Eds.) *Teaching and Learning by Doing Corpus Analysis. Papers from the Fourth International Conference on Teaching and Learning Corpora, Graz, 19/24 July 2000*. Amsterdam/Atlanta GA: Rodopi, 27-41.
Toolan, M. (2002). *Critical Discourse Analysis: Critical concepts in linguistics* (Vol. I-IV). London: Routledge.
van Dijk, T. (Ed.) (2011). *Discourse Studies. A multidisciplinary introduction*. London: Sage.
Wodak, R. & Meyer, M. (2009). *Methods for Critical Discourse Analysis* (2nd ed.). London: Sage.
Wood, S.N. (2006). *Generalized Additive Models: an introduction with R.* Boca Raton: Chapman and Hall/CRC Press

[1] However, the technique can also be used when the focus is the diachronic frequency development of a linguistic feature.
[2] The terms 'statistically significant' and 'statistical significance' will be used interchangeably with 'significant' and 'significance'.
[3] See Section 5.1 for a discussion of problems associated with linking troughs with contextual elements.
[4] The study presented in this paper formed part of the ESRC funded project "The representation of Islam and Muslims in the UK press, 1998-2008", ESRC reference RES-000-22-3536. For details, see http://www.ling.lancs.ac.uk/activities/826/. Note that, despite the title, the corpus comprises articles up to the end of 2009 (i.e. it spans 12 years).
[5] For detailed overviews of trends and approaches in critical discourse studies, see Toolan (2002), van Dijk (2011), Wodak & Meyer (2009).
[6] http://www.nytimes.com/interactive/2012/01/24/us/politics/0124-words.html?ref=politics (figures), http://www.nytimes.com/2012/01/25/us/politics/state-of-the-union-2012.html?_r=1 (article).
[7] For a detailed discussion of corpus-based approaches to (critical) discourse studies, see Baker (2006).
[8] For surveys of corpus-based diachronic studies, see Mair & Leech (2006), McEnery & Gabrielatos (2006: 54-56), McEnery et al. (2006: 178-194) and McEnery & Hardie (2012: 94-118). For diachronic studies within CADS (Corpus-assisted Discourse Studies), see Partington (2010).
[9] For example, the diachronic part of the Helsinki Corpus spans almost a millennium: 850–1710 (Kytö 1996).
[10] The actual reference time-unit is not important; what is important is that it is used consistently in all measurements. Of course, the selection of different time units will result in different sets of granularity scores. However, this is immaterial, as the scores are to be used comparatively. For example, if the year is used as the time-unit, then a diachronic corpus study using yearly sampling points will have a granularity score of 1. A study using half the sampling points (i.e. a sampling point every two years) will have a score of 0.5, whereas a study using twice the sampling points (i.e. a sampling point every six months) will have a score of 2.

[11] The searchable database we used to collect the articles, Nexis UK, categorised most newspapers as incorporating their Sunday editions. This was not the case for *The Observer*, which is sometimes viewed as a Sunday version of *The Guardian*. Additionally, *The People* is a Sunday-only newspaper which has no daily equivalent and *The Business* is a weekly newspaper which relaunched as a magazine in 2006 and then closed early in 2008. In our analysis, we have preserved these categorisations from Nexis UK, resulting in 12 newspapers.

[12] The selection of 1998 as the earliest period in the corpus was dictated by the duration and funding of the project.

[13] Please note that the symbol "!" is used as a wildcard in the database query system (e.g. *islam!* would return the forms *Islam, Islamic, Islamist, Islamists*).

[14] That is, the corpus frequency of *Moslem(s)* is a mere 4.6% of that of *Muslim(s)*.

[15] The term 'brachychrony' refers to time-spans which are deemed too short to merit the term 'diachrony', but long enough to allow for linguistic changes to be registered, and, thus, for the term 'synchrony' to be deemed unsuitable.

[16] For all papers reporting on different aspects of the project, see http://ucrel.lancs.ac.uk/projects/rasim/

[17] We accessed Google News at , http://news.google.com/

[18] The main link to Islam is that the bombing was carried out just before Ramadan.

[19] Jack Straw, member of parliament, leader of the House of Commons and ex-foreign secretary, said in a newspaper article that the veil is a "visible statement of separation and of difference" and that the practice of wearing it would make community relations in the UK harder.

[20] The Archbishop of Canterbury, Dr. Rowan Williams, in a speech, supported the introduction of sharia law in the UK, because it could support social cohesion.

[21] The straight line in this figure is a regression line. However, its R-squared value (coefficient of determination) is extremely low (0.1667, with 1 being the optimum value), so it does not reliably reflect fluctuations and diachronic trends in the data: only about 17% of the fluctuations in the monthly average number of articles can be explained by the different dates (monthly period).

[22] It must be clarified that, in the analysis reported in this section, major peaks were identified impressionistically (i.e. they were visually prominent).

[23] It does not seem unreasonable to expect this observation to apply to all types of periodicals.

[24] See www.r-project.org.

[25] Note that we chose a month by month analysis. Another important decision facing analysts using such techniques is to select the length of time that each sample point represents. We chose to take each month as our sample point as we were interested in studying long term trends – we did not want to look at daily peaks, for example, nor did we want to look at trends on a year to year basis. There may be very good reasons for altering the time span covered by each sample point, however, depending on the research question being pursued.

[26] The only exception is the *Business*, which shows a decrease.