Establishing a ‘new normal’: detecting fluctuating trends in word frequency over time

Matt Gee, Andrew Kehoe, and Antoinette Renouf

Birmingham City University, UK

Abstract

In this chapter we introduce statistical methods and associated visualisations for the analysis of lexical change on a monthly basis in a 1.8-billion word news corpus spanning over 30 years. In previous work (Kehoe et al. 2022) we found examples of word frequency change in a data-driven manner by applying existing statistical tests. An ongoing limitation is that, as our diachronic corpus grows, so too does the possibility of a word exhibiting multiple frequency changes in different directions. This chapter reframes the problem as one of time-series segmentation, dividing the frequency history of a word into timespans exhibiting consistent upward or downward change. We then determine reasons for such changes by applying horizon graph visualisations to collocates.

Keywords: lexical change, collocation, time series, visualisation, diachronic

1 Introduction and background

In this chapter we conduct a diachronic study of a corpus covering over 30 years of mainstream UK news text by developing new methods for analysing time series data and visualising collocational change over time. Such time series analysis has been of growing interest in the study of language change, supported by resources such as *Google Books N-grams* (Michel et al. 2011), the *News-On-the-Web* corpus (Davies 2013), and, more generally, Twitter (created in 2006) and various text archives (e.g. *UK Web Archive* and *LexisNexis* news archive). These provide time-indexed textual data, at sufficient scale and granularity that language change can be studied at monthly or even daily intervals with minimal issues of data sparsity. With such granularity comes the possibility of constructing time series of word frequency, which in turn opens up novel analytical techniques in linguistic study.

Data sources of this nature have received considerable attention under the umbrella of 'culturomics', in particular the *Google Books* dataset, consisting of over 5 million digitised books with year-indexed word and n-gram frequencies from 1800 onwards. Since the *Google Books N-grams* were released (Michel et al. 2011), several studies have made use of them to investigate cultural and societal change through the lens of word frequency (e.g. Gao et al. 2012, Greenfield 2013). However, concerns have been raised

over the use of approaches based solely on word frequency in drawing conclusions about cultural and societal change (see Hilpert 2020 for a discussion), and the lack of consideration of the linguistic context in which frequency change occurs is a major limitation. Importantly for the purposes of this chapter, most previous work in the field of culturomics has been limited by the fact that authors had to select specific words for investigation, thus relying on intuitions about language use as a first step.

Other studies have attempted to discover frequency change over time in a bottom-up, data-driven manner, albeit with a much narrower scope than our 30-year analysis. For example, Grieve et al. (2017) developed methods to find words consistently increasing in frequency within one year of Twitter data. Twenty-nine words were discovered and subjected to further analysis, showing that those words that do increase will either quickly fall out of usage (e.g. *bruuh*) or gradually stabilise in frequency (e.g. *fleek*). In our previous work (Kehoe et al. 2022) we presented approaches for finding word frequency change in a data-driven manner that were based on methods we first developed 20 years earlier. We investigated three types of frequency change in time series: long-running trends, in which a word increases or decreases in frequency over the course of multiple years; seasonal variation, in which changes in frequency are observed during the same months each year; and sudden upward jumps in frequency for otherwise rare words.

Change in the collocational profiling of a word has seen increased attention in Corpus Linguistics. Brezina (2018: 245-246) developed an

approach called Usage Fluctuation Analysis (UFA) to study language change by finding points in time exhibiting changes in the collocational profile of a word. By measuring the overlap in collocates between adjacent time points (e.g. between adjacent years), UFA indicates when the usage of the node word has changed. This approach has been applied to the study of discourse in historical corpora (McEnery et al. 2019) and parliamentary debates (Baker et al. 2017). Studies in Distributional Semantics have also concerned themselves with word meaning change, including a recent SemEval challenge (Schlechtweg et al. 2020). For example, Tsakalidis et al. (2019) used data from the *UK Web Archive* to find words changing in meaning between 2000 and 2013 with a method based on word embeddings (vector representations of words capturing contextual information). However, whilst the focus of these studies was on detecting usage/meaning change, the concern of this chapter is first to identify word frequency change and then to employ collocational methods to assist in interpretation. In Kehoe and Gee (2009) we developed a visual approach to studying collocational change, which can assist in studies where many thousands of examples would otherwise need to be inspected when trying to determine possible reasons for word frequency change.

The goal of the present chapter is to re-evaluate and refine our methods, resulting in new approaches for studying corpus-derived time series data. This section continues with a description of our previous methods and remaining issues we wish to address, and Section 2 describes the new

approaches we have developed to resolve these issues. Within each section, we first discuss the automatic detection of frequency change and then use collocational information to assist with the subsequent semi-automated analysis. Then, in Section 3, we test the methods and visualisations with multiple words. Finally, we end the chapter with conclusions and recommendations.

1.1 From corpus to time series

The corpus used in this chapter represents over three decades of UK newspaper reporting and has featured in several previous studies of language change (Kehoe and Gee 2009, Renouf 2013, Renouf and Kehoe 2013, Renouf 2018, Kehoe et al. 2022). We acknowledge that the corpus represents a single text-type, so our findings may not be generalisable to language change as a whole. However, using newspaper text has some important advantages. Firstly, the texts can be dated with very high accuracy as each article includes a date of publication. Secondly, newspapers publish a consistent stream of articles over long periods, allowing a large corpus to be constructed whilst avoiding issues of data sparsity. Our corpus contains articles published in the *Independent* and *Guardian* newspapers between January 1989 and December 2020, totalling 2.9 million texts and 1.8 billion tokens. More specifically, the *Independent* data is sourced from plain-text CD-ROM archives covering 1989-1999, totalling 396 million tokens, whereas 2000 onwards consists of

articles from the *Guardian* website, totalling 1.4 billion tokens (see Kehoe et al. 2022 for a full breakdown).

To enable diachronic study, the corpus is split into monthly sub-corpora (384 between 1989 and 2020 inclusive). For each month we generate word frequency information which is stored in a bespoke database.[1] Looking up a word in the database returns time series data (a combination of date and word frequency) for that word over all months of the corpus (23 occurrences in January 1989, 19 occurrences in February 1989, and so on). Subsequently, the frequencies are normalised against the size of each month of the corpus such that the word frequencies are transformed into frequencies per million words on a month-by-month basis. This is essential as there are fluctuations in the number of articles and tokens in each month, caused by the number of articles made available online increasing in recent years and differences in the number of days per month. A visual example is shown in Figure 1: a time series of frequency per million words for the word *viral* in each month of the corpus, along with a line showing smoothing over a 36-month window.[2]

[1] In this study we focus on word forms. However, the system is also capable of processing lemmatised and POS-tagged data.
[2] Specifically, the smoothing is calculated using a KZ (Kolmogorov–Zurbenko) filter (see Yang & Zurbenko 2010 for a description), with parameters span = 10 and iterations = 3.

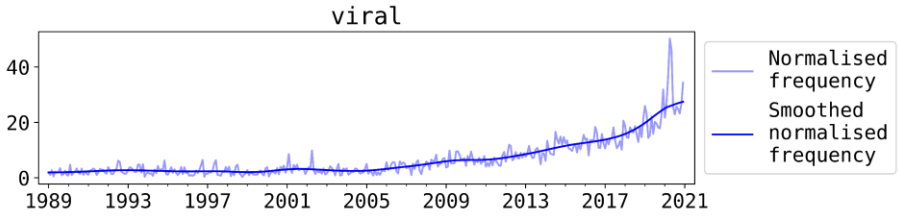

Figure 1. Frequency per million words of viral *and smoothed frequency.*

1.2 Identifying changes in frequency

In our previous work (Kehoe et al. 2022), we used three statistical tests which could be applied to a word's time series to determine the presence of three types of pattern: Cox's sequential test (Cox 1952), to find long-running trends; shifts in mean between adjacent windows of 12 and 24 months, to find sudden frequency jumps; and the coefficient of variation measured over monthly means, to find seasonal patterns. In that chapter we described our success in uncovering words which exhibit these patterns, but noted that trends can be obscured when they occur in opposite directions in quick succession. To be able to detect a downward trend following a pronounced upward trend (or vice versa), we must find an appropriate way of resetting our statistical tests, or establishing a 'new normal' from which subsequent variations can be found. In addition, our test for sudden changes was too sensitive to outliers and simultaneous trends (producing false positives), limiting our investigation to types that exhibited the most extreme changes. Being able to accurately pinpoint the beginning and end of a trend and points

of sudden change is useful for subsequent stages of analysis, such as selecting concordance examples from the trend period or automating part of the analysis by investigating collocational change.

The bottom sub-plot of Figure 2 shows the results of Cox's sequential test. The test determines the extent to which frequency is correlated with time and determines upper and lower confidence intervals which, if crossed, indicate an upward or downward trend. The word *viral* shows an upward trend from 2000 onwards. The test crosses the upper confidence threshold in 2008 and stays above the threshold until the end of the corpus (indicating that the frequency continues to increase). Analysing a simple trend pattern like this is straightforward, but the trend test as originally formulated in 2000 was calculated using the entire history of our news corpus which, at that point, was only 10 years. As our corpus has grown to cover 32 years, so too has the possibility of a word exhibiting multiple trends across the corpus. Reducing the window size for the test provides some improvement, but, as described below, the test result still lags behind the frequency data, making it difficult to find precise periods in which trends occur.

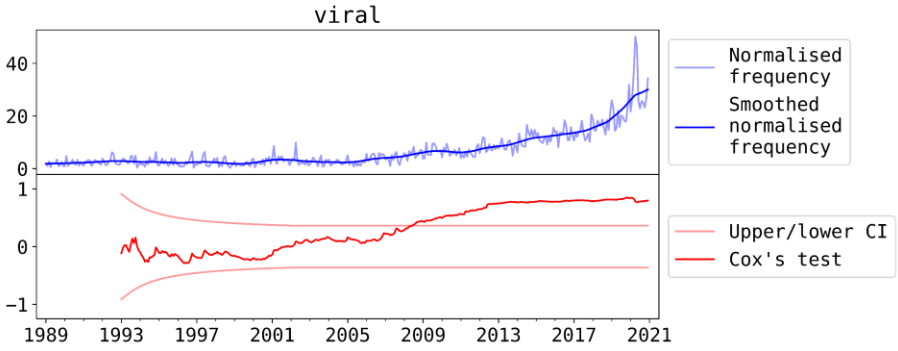

Figure 2. Top: Frequency per million words of viral *and smoothed frequency. Bottom: The result of Cox's sequential test, with confidence intervals at the 99.9% level.*

Such an example is shown in Figure 3, which exhibits multiple upward and downward trends in the frequency of the word *public*, each running over several years. It is clear from the top sub-plot in Figure 3 that *public* is decreasing in frequency between 2002 and 2007, but the test results in the bottom sub-plot indicate an upward trend in frequency during this period due to a previous upward trend between 2000 and 2002. This happens because the test requires an historical period of consistent change to discover a trend.[3]

Instead of attempting to detect a trend and then determine where it started and ended, our new solution in this chapter is to split the time series into periods of consistent upward or downward change and then test each period to determine if the trend is noteworthy. This process allows us to pinpoint the start and end points of trends more accurately. We combine this with a test to find abrupt shifts in a time series (Boulton & Lenton 2019), which the authors claim is robust to noise and simultaneous trends. Thus, we

[3] It should be noted that the position at which Cox's sequential test crosses the confidence intervals is dependent on both the duration and the extent of the frequency change.

can also identify sudden jumps or falls in frequency that occur over short periods. We describe the process in detail in Section 2.

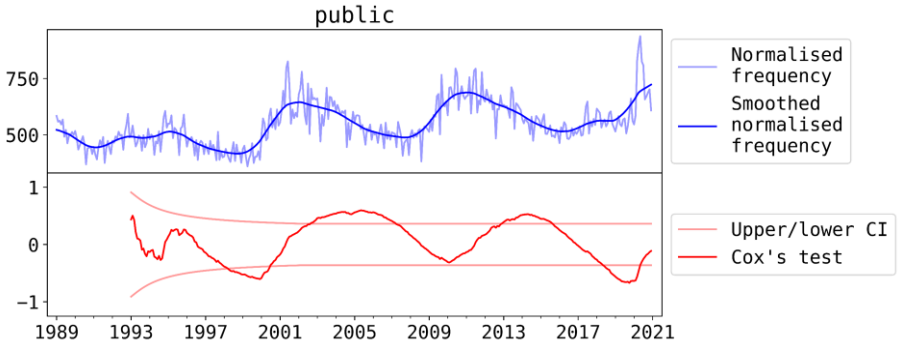

Figure 3. Top: Frequency per million words of public *and smoothed frequency. Bottom: The result of Cox's sequential test with confidence intervals at the 99.9% level.*

1.3 Exploring collocational change

Some fluctuations in word frequency are easy to explain. For example, the cyclical patterns observed for *Christmas* and *Olympics* are determined by real-world events, and the upward and downward frequency changes for *Instagram* and *iPad* relate to technological changes. However, the reasons for other changes can be harder to determine, especially with high frequency words where many thousands of examples need to be inspected. To assist in explaining such changes, we adopt a collocational approach.

In Kehoe and Gee (2009) we developed a visualisation to inspect collocational change which plotted collocational scores over time as a heat map. This was used to investigate collocational changes associated with the

financial crisis of 2008. The collocate heat map for the node word *credit* (Figure 4) shows, for example, that a popular term used in the news at the time, *credit crunch*, was not a new coinage but had been used previously in 1992 and 1999. We believe this method warrants further development. In the original paper, the z-score measure was used to calculate monthly collocate scores, which were then converted to values on a colour scale ranging from light to dark red based on low to high collocation scores. These colours were plotted on a heat map with dates on the x-axis and collocates on the y-axis. Each month of each collocate was drawn as a coloured block in a shade of red. While a minimum score to plot (the lightest red in the colour scale) could be established based on significance thresholds, the choice of a maximum (the darkest red) was somewhat arbitrary. This is noticeable in Figure 4 with the collocates *card*, *Suisse*, *cards*, *rating*, and *consumer*, which are displayed with the darkest colour throughout, thus providing no indication of whether any change is happening for those words. On reflection, we also have concerns over the validity of comparing z-scores across differently sized monthly sub-corpora. The connection between sample size and the z-score is such that scores are higher for larger samples when all other frequencies remain in proportion. In Section 2 below we describe an alternative approach to such a visualisation which addresses these issues.

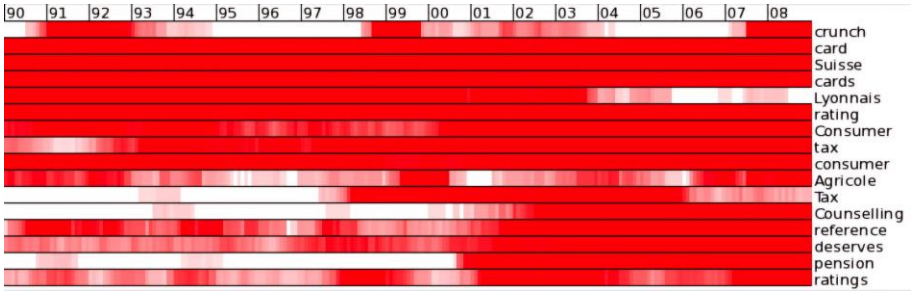

Figure 4. Heatmap for the top 16 collocates of credit.

2 Method

As noted above, our goal is to identify abrupt shifts and long-running trends in time series data. To find abrupt shifts, we turn to the approach described by Boulton and Lenton (2019). Our initial experiments indicated that seasonal variation can still affect this test, so we first apply time series decomposition to extract trend and seasonal components from the series (see Kehoe et al. 2022). Our approach to finding long-running trends is a refinement on the process discussed above, which now includes pre-processing steps to identify potential trend periods (where frequency change progresses in the same direction) before applying a statistical test (again using Cox's sequential test). We first apply the test for abrupt shifts, and then test for long-running trends between the identified shifts. As a result, the time series is split into segments, where each segment represents a direction and type of change (upward/downward abrupt shift, upward/downward trend, or

no change). Our refined approach is as follows, with the steps shown in Figure 5 for the word *recovery*):

1. Prepare the time series for the type as frequency per million words per month. An initial fit of the trend component is shown using a KZ (Kolmogorov–Zurbenko) filter (see Yang & Zurbenko 2010). This is equivalent to the smoothed frequency in Figure 1.

2. Decompose the time series to extract the seasonal component (calculated as monthly means) and divide by this to reduce monthly variation in the series.

3. Identify abrupt shifts in frequency using the *asdetect* algorithm (Boulton & Lenton 2019), which tests for sequences of extreme gradient in short windows, specifically:

    a. calculate the gradient for all windows (of size 11) across time (using linear regression);

    b. calculate the median and MAD (median absolute deviation) of the gradients;

    c. if the difference between the gradient in each window and the median is more than three times larger than the MAD, then apply a 'yes vote' for each point in the window (votes are stored in a parallel series where each vote adds one divided by the window size to fit the range -1 to 1; this is shown in the third sub-plot);

    d. points with enough votes are considered points of abrupt shift based on values crossing a threshold (shown in the third sub-plot with

horizontal lines for the threshold and vertical lines for the abrupt shift detection points).

4. Re-fit the trend component now abrupt shifts have been identified. Segment the series at peaks and troughs of the trend component. Four additional segment boundaries are found, as shown by vertical lines in the fourth sub-plot.

5. Test the trend in each resulting segment to filter out minor changes using Cox's sequential test. The fifth sub-plot shows the five segments determined to exhibit trends of sufficient scale, which are shown by the trend component being drawn in green (for upward trends) or red (for downward trends).

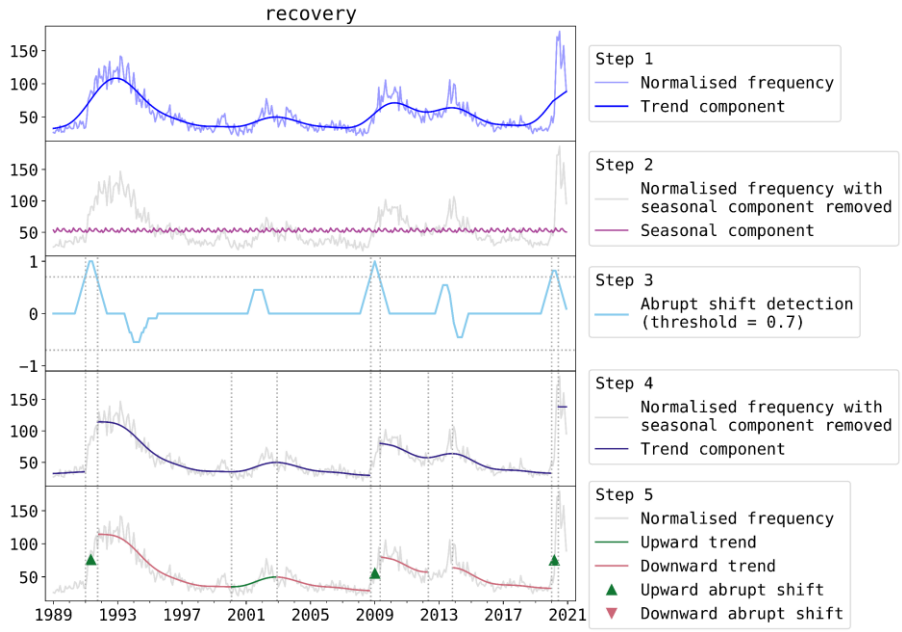

Figure 5. The five steps of the method to find frequency change segments. Sub-plots: 1. Frequency per million words for recovery*, with smoothed frequency; 2. Result of decomposition and removing seasonal variation; 3. Result of applying the test for abrupt shifts; 4. Result of segmenting at peaks and troughs of the trend; 5. Result of the process showing abrupt shifts with triangle markers and the trend line where a trend was found.*

Figure 5 shows the result of the process for the word *recovery*. The third sub-plot shows the results of the abrupt shift detection, which outputs values between -1 and 1 and crosses the threshold of ±0.7 (also used by Boulton & Lenton 2019 in their testing) in three places. The fifth sub-plot shows the abrupt shifts with triangular markers in the centre point of the abrupt shift (upward in May 1991, January 2009, and March 2020) and shows long-running trends by drawing the trend component (downward November 1991 to January 2000, upward February 2000 to November 2002, downward December 2002 to September 2008, downward June 2009 to April 2012, downward November 2013 to December 2019). InSection 3 we test the process with several words and discuss the results.

2.1 Visualising collocational change

We investigate collocational change using visualisations, which we see as a useful tool to further explore and potentially explain frequency change after a trend has been discovered. We take a span-based approach, where collocation is defined as the co-occurrence of a pair of words within a fixed span. To study this phenomenon over time, the year-by-year word co-occurrence frequencies are computed over the 32-year period of the corpus and stored in a database enabling lookup by node word. We test our method with a span of 4 words (left and right of a node word), which we consider a

useful starting point for exploratory corpus work in English whilst producing databases of manageable scale (see Collier 1993), but accept this may need to be adapted on a case-by-case basis. For a node word, the collocates to visualise are selected based on their collocational strength in the whole corpus (the 100 top collocates based on the z-score measure). Then, the co-occurrence frequency data for the selected collocates is converted to a proportion of the node frequency, providing a collocate frequency that is normalised with respect to the yearly node word frequency. For example, in the year 1998 *credit* occurs 3,198 times and *crunch* co-occurs with *credit* 75 times, resulting in a proportion of 0.023. The yearly results are presented using a visualisation tool known as a horizon chart (Saito et al. 2005).

A horizon chart is a space-efficient way of displaying a series that combines an area chart with colour information similar to a heat map. Horizon charts can show greater precision in a small vertical space than a line chart, making it possible to draw many sub-plots one above the other when they share a common x-axis. Each plot can be thought of an area chart (filled line chart) which has been sliced horizontally into equal-sized bands. Figure 6 shows a simple example of this, in which five bands of the chart are coloured in progressively darker shades from bottom to top and then layered on top of each other to create a single row of the horizon chart. The bottom slice is given a shade that represents the lowest values. The next slice is given a shade that represents higher values and is superimposed on top of the bottom slice so that both occupy the same space in the chart. This layering

continues for the remaining slices. Figure 7 shows the result for the word *credit* with plots for the top 10 collocates.

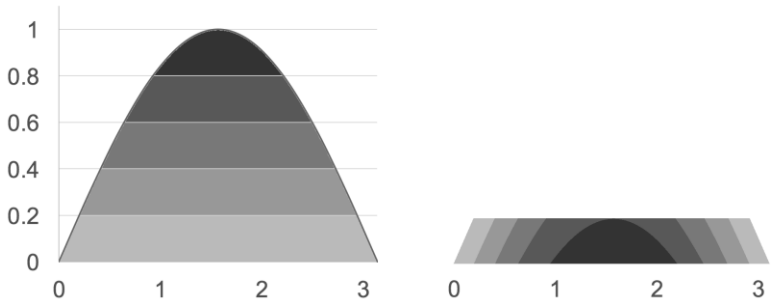

Figure 6. A simple example of a horizon chart. Left: the original area chart. Right: the result of layering the horizon slices where darker shades represent higher values.

To assist with interpretability, we follow suggestions from previous studies in visualisation. Heer et al. (2009) tested the number of bands with human analysts and recommended using no more than three for easy interpretation. The colour scale we apply is the Blues colour scale from ColorBrewer (Harrower & Brewer 2003), which is tool used to produce accessible and user-friendly colour schemes for data visualisation. In addition, we use yearly rather than monthly steps for several reasons: firstly, we are primarily interested in trends (rather than monthly variation); secondly, this reduces the processing and data storage requirements; thirdly, a balance must be made between granularity and co-occurrence frequency and, even at the billion-word scale of our corpus, we witnessed too much variation for useful interpretation when using monthly time steps. Finally, to show more detail per collocate, the maximum values plotted are determined on a per collocate

basis, i.e. the maximum value for *crunch* (0.359 in 2008, Figure 7) is different to the maximum value for *card* (0.240 in 2006), yet they are plotted with the same height and shade.[4] This can be configured to show a global maximum instead, but if some collocates are much more frequent than others it will be difficult to see detail in the changes for lower frequency collocates.

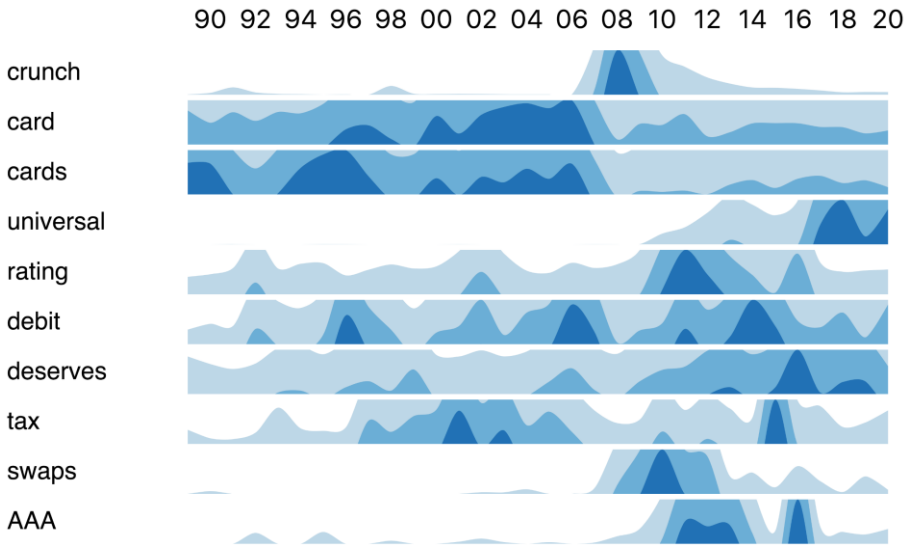

Figure 7. Horizon chart for the top 10 collocates of credit.

3 Results and Discussion

As a first step in our analysis, we ran the tests described in the previous section to determine how many trends or abrupt shifts in frequency we were able to detect for each of the words in our newspaper corpus across time.

[4] The minimum value is 0.

Taking all word types with a frequency per million words of one or greater gives us an initial list of 36,814 types, which we then restrict further by excluding grammatical words and numbers. This leaves us with the 23,435 types shown in Table 1, where the key finding is that over half the word types in our corpus (56.262%) do not exhibit any noteworthy trends or abrupt shifts during the time period covered. However, we are able to detect one trend or shift for almost a quarter of types (23.072%), with a further 11.905% exhibiting two trends or shifts and 5.291% exhibiting three. A small number of types exhibit up to nine trends or shifts during the period covered by the corpus. In total, we identified 8,608 trend periods and 2,742 abrupt shifts across the 23,435 types tested.

Table 1. Identification of trends and abrupt shifts in type frequency over time.

No. of trends/ shifts identified	No. of types	% of types
0	13,185	56.262
1	5,407	23.072
2	2,790	11.905
3	1,240	5.291
4	490	2.091
5	203	0.866
6	75	0.320
7	30	0.128
8	13	0.055
9	2	0.009
TOTAL	**23,435**	**100.000**

Before exploring examples of types for which we are able to detect trends or abrupt shifts in frequency over time, we look in more detail at those types for which we detect none. Table 2 lists the most frequent types in this category,

where the frequency shown is the monthly average (mean) per million words (pmw) across all months in the corpus.

Table 2. Top 20 most frequent types for which no trends or abrupt shifts are detected.

Type	**Mean freq pmw**		**Word**	**Mean freq pmw**
next	615.65		*international*	244.95
end	523.09		*campaign*	229.82
set	480.97		*tax*	224.97
put	461.34		*death*	223.24
four	443.23		*record*	211.31
already	372.65		*election*	203.48
high	330.39		*leader*	202.26
almost	324.97		*late*	201.68
taken	311.77		*several*	197.64
move	254.35		*try*	196.88

A chart for the most frequent word in Table 2, *next*, is shown in Figure 8. Although there is some fluctuation in the frequency of *next* over time, no upward or downward trends were identified (meaning that no trend lines are shown in Figure 8) and no abrupt shifts in frequency either (no upward or downward triangle markers). In fact, we find that the word *next* is seasonal, with a regular increase in frequency between September and December each year, usually peaking in November. This reflects the fact that newspaper articles make increased reference to *next year* towards the end of each year, as in example (1) from November 2008:

(1) The merger is now likely to be completed **next** year as the two sides continue to negotiate the terms. (2008-11)

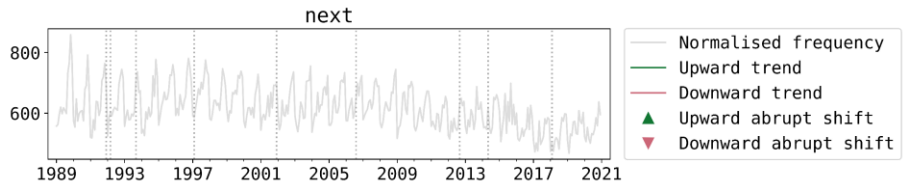

Figure 8. Frequency pmw of next.

Although our aim in this chapter is to detect trends and abrupt shifts in word frequency, it is actually reassuring to find that the majority of word types in our corpus do not exhibit such trends/shifts and remain fairly stable across the 32-year period. Table 2 includes a wide range of words, both in terms of topic coverage and word class, and it also seems to offer evidence to support the old adage that nothing is certain except *death* and *tax*(es), with no trends or abrupt shifts detectable in the frequency of either word across the 32 years. Now that we have reviewed a sample of the stable word types, the remainder of our analysis focuses on those types which exhibit trends/shifts according to the strict criteria set out in the previous section.

3.1 Downward trends

Table 3 lists the types in our newspaper corpus for which we are able to detect the most prolonged downward trends in frequency. Where we detect an abrupt shift in addition to a downward trend, this is shown in parentheses. For example, the first type in Table 3, *chairman*, exhibits adownward trend for 370 of the 384 months covered by the corpus. This downward trend begins

in the second month of the corpus (February 1989) and continues until September 2000, before resuming in November 2001 and continuing until the end of the corpus in December 2020. To give another example, the downward trend in *uprising* occurs in a total of 324 months, split into four distinct periods, including an abrupt shift between December 2011 and October 2012.

Table 3. Most prolonged downward trends (with abrupt shifts in parentheses).

Type	**Months**	**Segments**
chairman	370	1989-02 to 2000-09; 2001-11 to 2020-12
considerable	353	1989-05 to 2009-02; 2011-06 to 2020-12
sale	349	1989-02 to 1994-06; 1995-08 to 2005-03; 2007-01 to 2020-12
slipped	342	1989-02 to 2004-02; 2007-08 to 2020-12
virtually	325	1991-01 to 2018-01
impression	325	1990-12 to 2001-12; 2005-01 to 2020-12
uprising	324	1989-02 to 2000-06; 2001-01 to 2010-09; (2011-12 to 2012-10); 2012-12 to 2017-10
aroused	317	1989-02 to 1999-06; 2000-03 to 2016-02
discreet	315	1994-10 to 2020-12
distinguished	314	1994-11 to 2020-12
respectable	313	1994-12 to 2020-12
engine	311	1989-02 to 2003-02; 2009-03 to 2020-12
bound	310	1995-03 to 2020-12
virtues	309	1994-09 to 2007-02; 2007-10 to 2020-12
mainly	305	1993-10 to 1999-10; 2000-04 to 2010-08; 2011-11 to 2020-09
merely	303	1994-08 to 1998-05; 1999-08 to 2020-12
computer	302	1995-02 to 2010-12; 2011-10 to 2020-12
policeman	301	1993-07 to 1997-12; 2000-04 to 2020-10
telephone	301	1993-08 to 2010-01; 2010-06 to 2018-12
television	298	1994-07 to 2005-04; 2005-11 to 2013-07; 2014-10 to 2020-12

In order to examine the downward trends in Table 3 in more detail, we examine randomly selected concordance lines for each type. As these types are all declining in frequency over time, we choose examples from the first

month of the corpus (January 1989) when the frequency of each type was highest. For space reasons we present one concordance line each for a selection of the declining types in examples (2) to (10).

(2) Attending yesterday's business conference, Giovanni Agnelli, the **chairman** of Fiat, expressed confidence that Britain would eventually come round. (1989-01)

(3) It is hard enough for a trained **policeman** to investigate complaints against the police force because officers will not give evidence against their colleagues. (1989-01)

(4) In modern days, he who holds the **television** control is the person in charge. (1989-01)

(5) The Government has announced four licences to operate national mobile **telephone** services, known as 'telepoint'. (1989-01)

(6) He said he always wanted to be a news-reader. But he was not born to sound **respectable** enough for that trade, and never lost his Merseyside accent. (1989-01)

(7) an outstanding scientist and personality. He came from a **distinguished** medical family, but was inclined towards Classics at school. (1989-01)

(8) Escalating prices have **virtually** stopped British galleries from competing with institutions abroad. (1989-01)

(9) It seems he had **merely** meant to give the chap a sharp tap on the shoulder with a club. (1989-01)

(10) The consumer alarm **aroused** by salmonella and listeria has focused attention on the inconsistency. (1989-01)

Concordance lines (2) and (3) are examples of the kind of gendered language that has been declining in the British press and society more widely since the 1980s. The word *chairman* has been replaced by *chairperson* or just *chair*, although those terms tend to be used to refer to the leader of a meeting or small organisation, whereas the leader of a large company like Fiat in (2) is more likely to be referred to using a term such as *CEO* nowadays. Meanwhile, the word *policeman* in (3) has been replaced by *police officer* in modern usage, although it is interesting to note that this 1989 example does use the term *officers* later in the same sentence to refer to members of the police force more generally (the assumption seems to be that the 'trained' officers responsible for conducting internal investigations were likely to be men at that time).

Examples (4) and (5) relate to technological changes since the late 1980s, with *TV* replacing *television* (and, more recently, the whole concept of television being replaced by online streaming), and phone now more commonly used than the longer form *telephone*, particularly in the context of *mobile*. In fact, *(tele)phone* is now often omitted altogether in terms such as *mobile services* and *mobile networks*. That is to say that *mobile* has become

a synonym of *(tele)phone* rather than a hyponym. Examples (6) and (7) reflect the general decrease in formality in British newspapers that we reported on in our previous work (Kehoe et al. 2022), where we found that the term *Mr* had been decreasing in frequency since 1989. Here, we find a similar decrease in both *respectable* and *distinguished*, both of which are used to refer to men in these examples and both of which are used in discussions of social class and family background. Both downward trends continue from late 1994 to the end of the corpus.

Concordance lines (8) and (9) include two adverbs declining in use in the British press over the past three decades: *virtually* and *merely*. The former means *practically, nearly*, or *almost* in the context of (8), and the latter means *simply, only* or *just* in the context of (9). Both sound rather old fashioned to the modern reader, which is reinforced by the use of *chap* in (9). Another reason for the decline of *virtually* in this usage appears to be the more recent growth of *virtually* meaning 'by means of virtual reality' or simply 'on the internet' (cf. the slightly ambiguous example from March 2012 cited by Renouf (2017: 44): 'you could virtually travel there in advance using the app to set up meetings'). The growth of this new meaning seems to be reflected in the fact that the longer-term downward trend in the overall frequency of *virtually* that began in January 1991 stops after January 2018, whereas the downward trend in *merely* continues until the end of the corpus (Table 3). Finally, example (10) uses the word *aroused*, the frequency of which has declined in non-sexual contexts since the late 1980s, except in the specific

phrase *aroused suspicion*. A 2020s equivalent of example (10) would perhaps replace *aroused* with *caused, created* or *sparked*.

3.2 Upward trends

Table 4 lists the types in our corpus with the longest prolonged upward trends in frequency, shown in the same format as the downward trends in Table 3. Here, the top type, *ongoing* exhibits the longest possible trend, beginning in the second month of the corpus and continuing unbroken until the end.

Table 4. Most prolonged upward trends (with abrupt shifts in parentheses).

Type	**Months**	**Segments**
ongoing	383	1989-02 to 2020-12
telling	370	1989-02 to 2019-11
seeing	368	1989-02 to 2019-05; (2019-12 to 2020-03)
feels	367	1989-02 to 1998-09; 2000-02 to 2020-12
prioritised	366	1989-02 to 2005-11; 2006-09 to 2019-09; (2019-10 to 2020-04)
decades	364	1989-02 to 2019-05
focused	362	1990-03 to 2005-10; 2006-07 to 2020-12
understands	361	1989-02 to 2001-02; 2001-10 to 2008-09; 2010-01 to 2020-12
incredible	357	1989-02 to 1999-07; 2001-10 to 2020-12
globally	355	1989-02 to 2001-03; 2002-10 to 2009-08; 2009-11 to 2019-11; (2019-12 to 2020-04)
concerns	355	1989-02 to 2003-10; 2006-03 to 2020-12
mindset	353	1991-08 to 2020-12
photo	348	1989-02 to 2008-06; 2009-06 to 2018-12
life-changing	348	1991-07 to 2008-07; 2009-01 to 2020-11
struggled	346	1989-02 to 1999-08; 2001-01 to 2008-01; 2009-11 to 2020-12
citing	345	1989-02 to 2006-05; 2008-06 to 2013-06; 2014-09 to 2020-12

spotlight	342	1989-08 to 2002-10; 2003-08 to 2018-10
timeframe	341	1992-08 to 2020-12
stop	340	1989-02 to 1994-11; 1996-08 to 2003-12; 2004-06 to 2010-04; 2010-10 to 2019-11
moments	336	1989-02 to 2001-01; 2003-04 to 2019-03

As with the downward trends, we illustrate the upward trends with randomly selected concordance lines from the corpus but this time we choose examples from the end of the corpus, when the frequencies of these types are at their highest, see (11) to (16).

(11) The search for the missing person is **ongoing** and all possibilities remain open until she is found. (2020-12)

(12) She **feels** psychologically exhausted, like every day is a torment. (2020-12)

(13) I'm a mother of a teen who has **struggled** with an anxiety disorder for several years. (2020-12)

(14) Loss of smell can be **life-changing**; it removes an important part of your sense of self. (2020-12)

(15) I choose the freeing **mindset** of forgiveness, which is so healing. (2020-12)

(16) There are so many **incredible** places to visit in Britain and being on a bike gives you a different perspective. (2020-12)

Concordance line (11) for the type *ongoing* illustrates the same phenomenon we reported on in our previous study (Kehoe et al. 2022), where we found a

prolonged downward trend in the frequency of the word *yesterday* in *The Guardian* (more frequent on a single day in January 1989 than in the whole month of December 2017). Our explanation for that change also partly accounts for this one: the shift from the reporting of 'yesterday's news' in print to the reporting of *ongoing* stories in real-time online. It was, of course, possible for events to be ongoing when they were reported upon in print newspapers in the 1980s and 1990s, but this possibility increased with the advent of online newspapers in the late 1990s and has accelerated with the growth of online-first and online-only publication since. The earlier reasons for the upward trend in *ongoing* as far back as 1989 require further analysis.

The types exemplified in (12) to (15) all reflect an increasing openness in British society since the 1980s to discussing issues relating to emotions and mental health. Referring back to Table 4, *feels* (12) has been on a prolonged upward trend since February 1989 except for a short period at the end of the 1990s, while *struggled* (13) exhibits three extended upward trends with short breaks, *life-changing* (14) shows a long-term upward trend from July 1991 with a five-month gap in 2008, and *mindset* (15) is on a prolonged upward trend from August 1991 until the end of the corpus in December 2020.

Finally, example (16) illustrates the prolonged upward trend in *incredible*, which began at the earliest possible point in February 1989 and continued to the end of the corpus with a gap in the late 1990s and early 2000s. The concordance line in (16) uses *incredible* as a positive evaluator

and, after examining other concordance lines from across the corpus, we believe that the increase in this particular usage is largely responsible for the overall increase in the frequency of the type. Example (16) can be contrasted with the earlier use of *incredible* to mean literally 'not credible', as illustrated in example (17) from January 1989.

(17) Godfrey Hodgson made the following **incredible** statement: So far as de facto desegregation in the South was concerned, King and his movement were pushing at an open door. (1989-01)

In the following section we test our new semi-automated method for identifying such changes in meaning across time through collocational analysis.

3.3 Multiple abrupt shifts

We now turn our attention to word types that exhibit multiple changes. When multiple abrupt shifts occur, they do so in greater numbers than trends, so we discuss them separately. Table 5 lists the types for which our methods identify the largest number of abrupt shifts during the period covered by our corpus. The time series analysis results for selected terms are included in Figure 9.

Table 5. Types with multiple abrupt shifts.

Number of abrupt shifts identified	**Types**
8	*antisemitic, bailout, blog, midterm, que*
7	*double-dip, eurozone, localism, pdf, playoffs, selfie, tbs, www*
6	*anag, austerity, bloggers, blogging, blogosphere, blogs, caucus, caucuses, click, color, cybersecurity, declarer, dotcom, low-carbon, multi-party, peace-keeping, please, podcast, randomized, redacted, talkboard, tea-timely, zero-hours*

For some of the types exhibiting multiple abrupt shifts, the reasons behind the changes are simple to establish. Many of them can be considered topical, such that the type increases in frequency suddenly as a topic becomes newsworthy and then reduces in frequency quickly when the topic is no longer newsworthy. Table 5 includes such examples as *bailout, midterm, double-dip, low-carbon, zero-hours*, and *austerity*. The time series analysis for *austerity* is shown at the top of Figure 9.

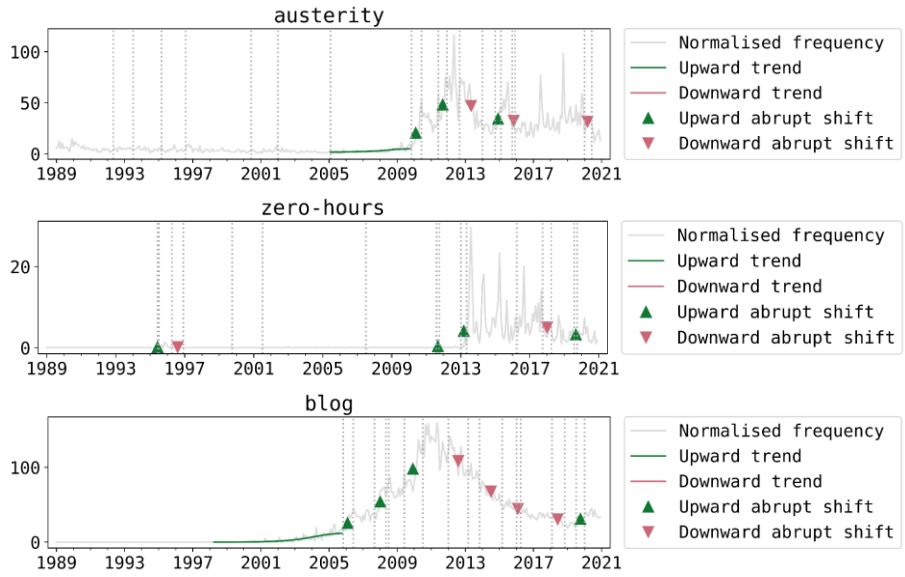

Figure 9. Results of our analysis process for austerity, zero-hours, *and* blog. *Abrupt shifts are marked by triangles. Where trends have been detected, the trend line is drawn.*

The upward abrupt shifts detected for *austerity* can be associated with newsworthy events. They are centred around February 2010 (the UK Government austerity programme was initiated in 2010) and September 2011, followed by a downward shift in May 2013, upward again in November 2015 (when the Treasury extended the austerity period), and downward in March 2020.

A similar pattern can be observed for *zero-hours* (Figure 9). The more recent upward abrupt shifts (September 2011, March 2013) concern discussions around working conditions and *zero-hours contracts*, as exemplified in (18). The topic remains in the news for some time, with *zero-hours* fluctuating in frequency but higher than it was previously, then the frequency reduces in January 2018. It is interesting to note the earlier upward

shift in July 1995, followed by a downward shift in August 1996, which concerns the same topic of *zero-hours contracts*. Interestingly, example (19) has the phrase in quotation marks and uses the word *whereby* to provide a gloss, indicating that the term was not widely known at that point. Although the frequency is much lower than the more recent increase, these changes are still (validly) detected as abrupt shifts. Also of note is the value in using time series data to view word frequency history, rather than snapshots at specific points in time which can miss such details (see Kehoe and Gee 2009, where we also noted similar use of quotation marks and glosses in early instances of the term *credit crunch*).

(18) If you want a job with guaranteed hours, to be forced on to a **zero-hours** contract denies you the psychological security that comes with knowing how much you will earn in a given month. (2013-12)

(19) The draft said that Labour would put an end to **'zero-hours** contracts', whereby workers are required to be on continuous call with no guarantee of work or pay. (1996-06)

Multiple abrupt shifts were also detected for types associated with newspaper features that ran for limited periods or with differing frequency between periods. As shown in example (20), this includes *tea-timely*, which is a pun repeated on pages announcing the availability of *The Guardian's* football

newsletter *The FIVER*. Similarly, abrupt shifts in the frequency of *please* are associated with formulaic messages directed towards readers (example 21), featuring in articles in the Media section of *The Guardian*.

(20) SIGN UP TO THE FIVER. Want your very own copy of **tea-timely** (ish) email sent direct to your inbox for free every weekday? (2009-12)

(21) For all other inquiries **please** call the main Guardian switchboard on 02072782332. If you are writing a comment for publication, **please** mark clearly "for publication". (2007-07)

Changes can also be observed that relate to orthographical aspects. For example, *tbs* is an alternative form of *tbsp*/*tablespoon* which is used for limited periods in recipes, and *color* is predominately used in the n-gram *of color* (67% of occurrences, e.g. *people of color*, *women of color*) with the US spelling (although the British spelling also occurs). We note that *color* occurs in *Guardian* articles across all sections of the newspaper, but mostly those reporting on American news (35% in the *US News* section) and articles by external contributors (22% in the *Comment is Free* section), as in examples (22) and (23) respectively.

(22) Together, that group would represent roughly 23 million Americans, disproportionately including women, people of **color**

and low-wage workers who make up the healthcare labor force. (2020-12)

(23) What I saw at this press conference was an inspiring tableau of today's America, four women of **color** offering moral and political guidance to a nation that is being goaded by its commander-in-chief to embrace the schoolyard taunts of a wanton bigot. (2019-07)

Another subset of abrupt shifts concern technological developments. Such examples from Table 5 include *blog, selfie, dotcom,* and *podcast*. The analysis results for *blog* are shown in Figure 9. The upward abrupt shifts centred around February 2006, January 2008, and December 2009 are of the kind we expect to be detected by the abrupt shifts test (i.e. the frequency rises quickly within a few months and a new level is established), but the multiple downward shifts detected from 2012 onwards may have been better matched as a long-running downward trend. This tends to occur for low frequency words where there is little change across the series as a whole, so that when change does occur it is deemed to be an abrupt shift.[5] We note the example of *recovery* in the next section (and discussed in Section 2) where a similar pattern is better matched by our method.

[5] i.e. the median absolute deviation of the gradients is very low, so when small changes occur, it seems this threshold is crossed easily.

3.4 Multiple trends

In the final part of our analysis we note the types which exhibit multiple trends according to our tests (shown in Table 6).

Table 6. Types for which multiple trends were detected.

Number of trends identified	**Types**
5	*fears, increased, public, recovery*
4	*actually, administration, ago, annoying, anxious, authority, became, better, born, budgets, came, collapse, council, councils, country, creative, customer, cutbacks, debts, department, effectively, following, good, got, governance, guy, hardline, interactive, just, key, little, market, media, national, nice, officer, okay, organisation, plans, probably, provide, quite, reports, resources, says, share, stop, struggling, support, surely, thing, unemployment, using, warned*

In comparison with the types with multiple abrupt shifts, the types exhibiting multiple trends are seemingly more general in terms of topic or meaning. Thus, we explore the context of their trends through collocation, using the horizon chart visualisation tool described in the Section 2.1. Figure 10 shows the time series segmentation analysis for three examples (*public, recovery, using*), which we discuss below.

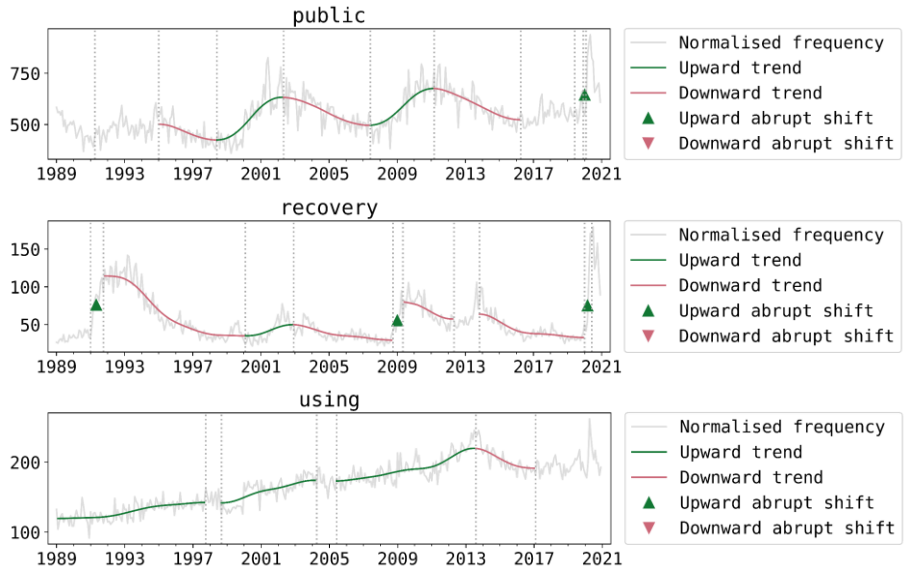

Figure 10. Results of our analysis process for public, recovery, *and* using. *Abrupt shifts are marked by triangles. Where trends have been detected, the trend line is drawn.*

The type *public* exhibits long running trends one after the other in opposite directions, from January 1995 (downward), June 1998 (upward), May 2002 (downward), June 2007 (upward), and finally March 2011 (downward until April 2016). The collocate horizon chart for *public* is shown in Figure 11. The relative frequency of the collocate *service* starts to increase from 1998 (with *services* and *service* peaking in 2001 and 2002 respectively), coinciding with the upward trend in the frequency of *public* during that period shown in Figure 10. Similarly, from 2008 to 2010, the collocates *sector*, *finances*, and *spending* increase in frequency, during another upward trend for *public*. These terms also decrease in collocate frequency as *public*'s frequency decreases during the subsequent downward trend. Thus, they help us to refine our view of the topics under discussion during those upward trend periods. In

fact, *public* is mostly a pre-modifier in these examples, with the collocates providing much of the topical information (*finances, spending, services*) and *public* specifying the domain in which the topics are relevant. Also of note in the graph for *public* in Figure 10 is the abrupt shift upwards in 2020, during which time the collocates *transport* and *health* increase, which, as can be seen from examples (24) and (25), relate to the COVID-19 pandemic.

(24) Burnham asked Matt Hancock to share all the information he has on Covid-19 in Greater Manchester to enable local **public** health teams to do their job, echoing similar concerns expressed by the mayor of Leicester. (2020-07)

(25) In a global pandemic, people across the world have avoided **public** transport systems where they can. (2020-11)

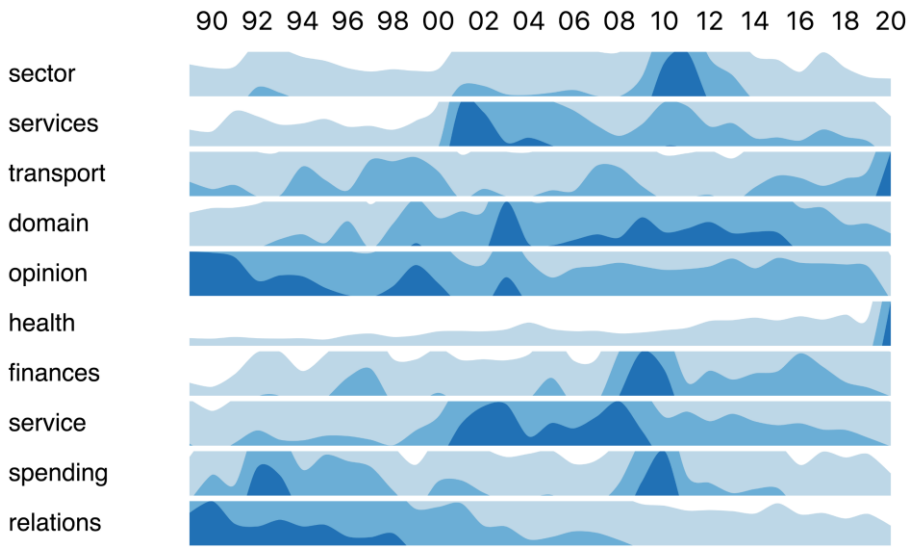

Figure 11. Horizon chart for the top 10 collocates of public.

Figure 12 shows the collocates of *recovery*, the strongest of which, *economic*, indicates a connection to the financial crises of 1991 and 2008 (as discussed in Kehoe and Gee 2009). During these periods, the frequency of *recovery* increases quickly (identified as upward abrupt shifts in Figure 10 centred around May 1991 and January 2009) and remains at an increased level for over a year. Collocates such as *economic, signs [of], fragile, slow,* and *sustained* also increase in frequency during the same periods, occurring in contexts related to finance (examples 26 and 27).

(26) Today's decision shows that the Bank feels more action is needed to kickstart growth, with economic **recovery** still fragile. (2009-08)

(27) They are about growth, growth, growth - creating the conditions for a sustained economic **recovery**, and an economic recovery is a function above all of confidence. (2010-10)

Subsequently, the frequency of *recovery* reduces gradually (identified as downward trends from November 1991 to February 2000 and from June 2009 to May 2012), with the collocates *economic* and *fragile* following similar reductions in frequency.

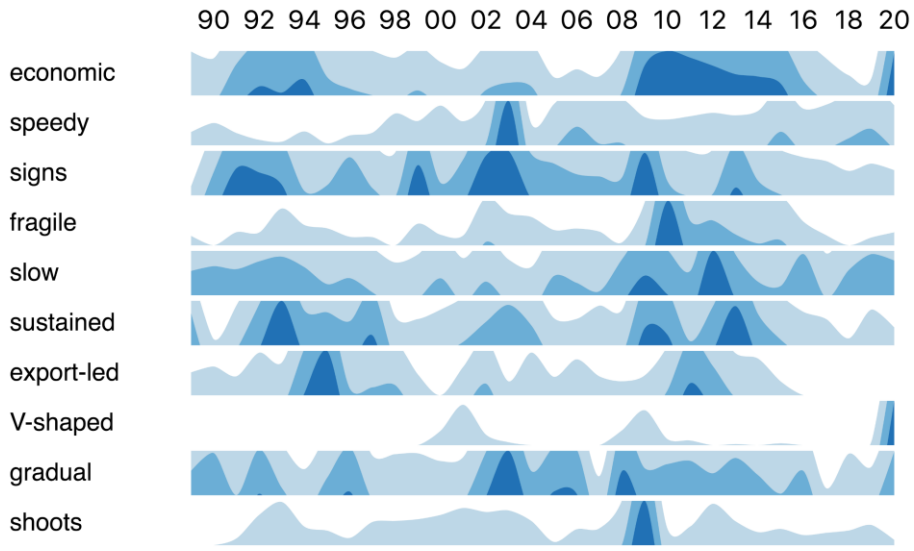

Figure 12. Horizon chart for the top 10 collocates of recovery.

The type *using* exhibits three upwards trends and one downward trend (Figure 10). The upward trends occur in the periods February 1989 to September 1997, September 1998 to March 2004, and June 2005 to July 2013. The downward trend is between September 2013 and January 2017. On visual inspection, the time series segmentation seemingly misses some subtle details, such as a dip between 1998 and 2000, immediately followed by a potential upward shift (in this case the test was far from crossing the threshold). Alternatively, *using* could be considered to exhibit a constant upward trend from 1989 to 2013, but our methods do not treat the frequency change in this way due to breaks in the trend (October 1997 to October 1998 and April 2004 to June 2005) that are a year or more in length. The highest ranked collocates (Figure 13) reveal that *using* occurs with words related to technology and the web, with *mobile* and *technology* increasing in

collocational frequency over similar periods to *using*. Platform-specific terms (*hashtag*, *Twitter*) start to collocate with *using* from 2008 onwards (two years after Twitter was launched) and peak at the same time as *using* (2012 to 2013). They are mentioned in articles to draw attention to *The Guardian's* own and others' Twitter content and when reporting on topics trending on social media, as shown in examples (28) and (29) from the period in question.

(28) The Guardian is inviting readers to share thoughts and memories of the Egyptian revolution on Twitter **using** the hashtag #jan25stories (2012-01)

(29) News of her arrest quickly spread online as Twitter users posted their outrage **using** the hashtag #daftarrest (2013-03)

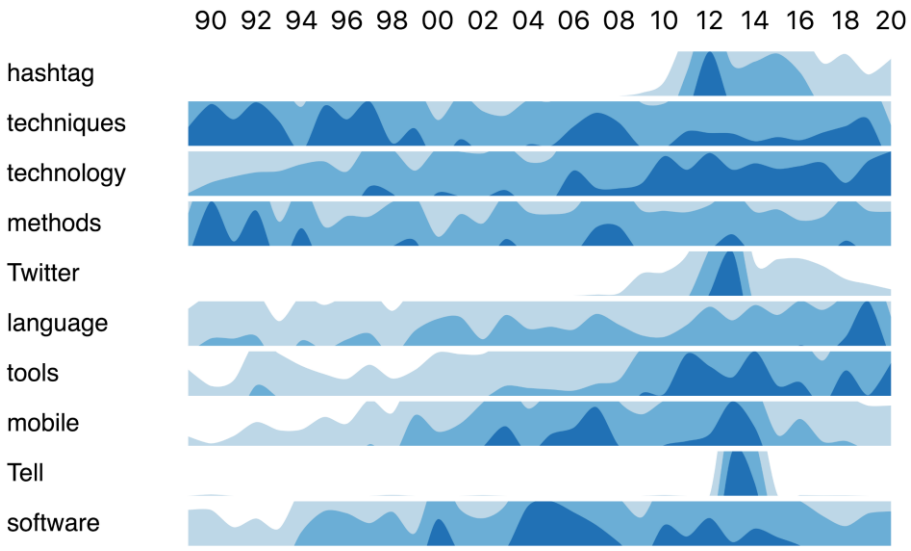

Figure 13. Horizon chart for the top 10 collocates of using.

4 Conclusion

In this chapter we set out to identify two types of word frequency change pattern in corpus-derived time series: abrupt shifts and long-running trends. Building on our previous work (Kehoe et al. 2022), we discussed the challenges involved in locating change points in a diachronic corpus, and tested new methods designed to resolve them. Overall, our results are encouraging. We used our approach to identify both changes that span the majority of our corpus and those that occur for limited periods only. We found the former to be associated with wider cultural and societal changes (*chairman, feels, life-changing*) and the immediacy of online news reporting (*ongoing*), whereas the latter are associated with technological developments and topical shift. We also presented refinements to our approach to visualising collocational change (Kehoe and Gee 2009), now using horizon charts. These charts proved useful in highlighting the reasons behind particular trends, such as the specification of topic domain for *public* and the association between *using* and technological advances. We reiterate the importance of investigating textual and collocational context when interpreting word frequency change in order to develop a full picture (cf. Kehoe and Gee 2009).

We must also note that our approach includes a number of steps which could be adapted on a case-by-case basis, including the choice of

decomposition models and fitting methods for the trend and seasonal components, of thresholds and window size for the *asdetect* algorithm, and of confidence levels for Cox's test. In this study, in which the method was tested on over 20,000 types, we made choices we thought applicable in the majority of cases. We discussed some cases above (*blog*, *using*) where tweaks to the method such as higher thresholds or longer smoothing windows could produce improved results. On the whole, we believe the methods explored in this chapter are an improvement on our previous approaches and hope that they will prove useful to others, with or without further refinement.

Regarding future work, we expect a move away from approaches which investigate frequency change in the first instance towards methods based on first categorising or clustering instances and subsequently investigating change for categories as a whole. Recent examples of such work include Murakami et al. (2017), Schneider (2022), and Clarke et al. (2022) who investigate changes over time in topic, meaning and discourse. There is potential to apply similar approaches to contextual spans of varying sizes (e.g. collocational, textual) and further investigate the types of change revealed.

References

Baker, Helen, Brezina, Vaclav & McEnery, Tony. 2017. Ireland in British parliamentary debates 1803–2005: plotting changes in discourse in a large volume of time-series corpus data. In *Exploring future paths for historical sociolinguistics* [Advances in Historical Sociolinguistics], Tanja Säily, Arja Nurmi, Minna Palander-Collin & Anita Auer (eds), 83–107. Amsterdam: John Benjamins.

Boulton, Chris & Lenton, Timothy. 2019. A new method for detecting abrupt shifts in time series. *F1000Research* 8: 746.

Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics*. Cambridge: Cambridge University Press.

Clarke, Isobelle, Brookes, Gavin & McEnery, Tony. 2022. Keywords through time: Tracking changes in press discourse of Islam. *International Journal of Corpus Linguistics* 27(4): 399–427. <https://doi.org/10.1075/ijcl.22011.cla>

Collier, Alex. 1993. Issues of large-scale collocational analysis. In *English Language Corpora: Design, Analysis and Exploitation*, Jan Aarts, Pieter de Haan & Nelleke Oosdijk (eds), 289–298. Amsterdam: Rodopi.

Cox, David, 1952. Sequential tests for composite hypotheses. *Mathematical Proceedings of the Cambridge Philosophical Society* 48(2): 290–299.

Davis, Mark. 2013. Corpus of News on the Web (NOW): 3+ billion words from 20 countries, updated every day, <https://www.english-corpora.org/now/> (18 July 2023).

Gao, Jianbo, Hu, Jing, Mao, Xiang & Perc, Matjaž. 2012. Culturomics meets random fractal theory: insights into long-range correlations of social and natural phenomena over the past two centuries. *Journal of The Royal Society Interface* 9: 1956–1964.

Greenfield, Patricia. 2013. The Changing Psychology of Culture From 1800 Through 2000. *Psychological Science* 24(9): 1722–1731.

Grieve, Jack, Nini, Andrea & Guo, Diansheng. 2017. Analyzing Lexical Emergence in Modern American English Online. *English Language and Linguistics* 21(1): 99–127.

Harrower, Mark & Brewer, Cynthia. 2003. ColorBrewer.org: an online tool for selecting colour schemes for maps. *The Cartographic Journal* 40(1): 27–37.

Heer, Jeffrey, Kong, Nicholas & Agrawala, Maneesh. 2009. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 1303–1312. New York: Association for Computing Machinery.

Hilpert, Martin. 2020. The Great Temptation: What Diachronic Corpora Do and Do Not Reveal about Social Change. In *Corpora and the Changing*

Society. Studies in the Evolution of English, Paula Rautionaho, Arja Nurmi & Juhani Klemola (eds), 3–27. Amsterdam: John Benjamins.

Kehoe, Andrew & Gee, Matt. 2009. Weaving Web data into a diachronic corpus patchwork. In *Corpus Linguistics: Refinements and Reassessments*, Antoinette Renouf & Andrew Kehoe (eds), 255–279. Amsterdam: Rodopi.

Kehoe, Andrew, Gee, Matt & Renouf, Antoinette. 2022. A data-driven approach to finding significant changes in language use through time series analysis. In *Broadening the Spectrum of Corpus Linguistics: New approaches to variability and change*, Susanne Flach & Martin Hilpert (eds), 285–318. Amsterdam: John Benjamins.

McEnery, Tony, Brezina, Vaclav & Baker, Helen. 2019. Usage Fluctuation Analysis: A new way of analysing shifts in historical discourse. *International Journal of Corpus Linguistics* 24(4): 413–444. <https://doi.org/10.1075/ijcl.18096.mce>

Michel, Jean-Baptiste, Yuan Kui Shen, Aiden, Aviva P., Veres, Adrian, Gray, Matthew K., The Google Books Team, Pickett, Joseph P., Hoiberg, Dale, Clancy, Dan, Norvig, Peter, Orwant, Jon, Pinker, Steven, Nowak, Martin A. & Lieberman Aiden, Erez. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* 331(6014): 176–182. <https://doi.org/10.1126/science.1199644>

Murakami, Akira, Thompson, Paul, Hunston, Susan & Vajn, Dominik. 2017. 'What is this corpus about?': using topic modelling to explore a

specialised corpus. *Corpora* 12(2): 243–277. <https://doi.org/10.3366/cor.2017.0118>

Renouf, Antoinette. 2013. A finer definition of neology in English: The life-cycle of a word. In *Corpus Perspectives on Patterns of Lexis*, Hilde Hasselgård, Jarle Ebeling & Signe Oksefjell Ebeling (eds), 177–208. Amsterdam/Philadelphia: John Benjamins.

Renouf, Antoinette. 2017. Some corpus-based observations on determinologisation. *Neologica* 11: 21–48. <https://doi.org/10.15122/isbn.978-2-406-06995-9.p.0021>

Renouf, Antoinette. 2018. Big Data: Opportunities and Challenges for English Corpus Linguistics. In *From Data to Evidence in English Language Research*, Carla Suhr, Terttu Nevalainen & Irma Taavitsainen, (eds), 27–65. Leiden: Brill.

Renouf, Antoinette & Kehoe, Andrew. 2013. Filling the gaps: Using the WebCorp Linguist's Search Engine to supplement existing text resources. *International Journal of Corpus Linguistics* 18(2): 167–198.

Saito, Takafumi, Miyamura, Hiroko Nakamura, Yamamoto, Mitsuyoshi, Saito, Hiroki, Hoshiya, Yuka & Kaseda, Takumi. 2005. Two-tone pseudo coloring: Compact visualization for one-dimensional data. In *IEEE Symposium on Information Visualization INFOVIS 2005*, 173–180. IEEE.

Schlechtweg, Dominik, McGillivray, Barbara, Hengchen, Simon, Dubossarsky, Haim & Tahmasebi, Nina. 2020. SemEval-2020 Task 1:

Unsupervised Lexical Semantic Change Detection. <https://doi.org/10.48550/arXiv.2007.11464>

Schneider, Gerold. 2022. Systematically Detecting Patterns of Social, Historical and Linguistic Change: The Framing of Poverty in Times of Poverty. *Transactions of the Philological Society* 120(3): 447–473. <https://doi.org/10.1111/1467-968X.12252>

Tsakalidis, Adam, Bazzi, Marya, Cucuringu, Mihai, Basile, Pierpaolo & McGillivray, Barbara. 2019. Mining the UK Web Archive for Semantic Change Detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 1212–1221. Varna: INCOMA Ltd. <https://doi.org/10.26615/978-954-452-056-4_139>

Yang, Wei & Zurbenko, Igor. 2010. Kolmogorov–Zurbenko filters. *WIREs Computational Statistics* 2(3): 340–351.

Acknowledgement

We thank the Engineering and Physical Science Research Council (EPSRC) for research grants: GR/L08243/01 (APRIL), GR/R16884/01 (WebCorp) and EP/E001300/1 (WebCorpLSE), which funded the infrastructure for this study.