Usage Fluctuation Analysis

A new way of analysing shifts in historical discourse

Tony McEnery, Vaclav Brezina and Helen Baker
Lancaster University

This article introduces a methodology for the diachronic analysis of large historical corpora, Usage Fluctuation Analysis (UFA). UFA looks at the fluctuation of the usage of a word as observed through collocation. It presupposes neither a commitment to a specific semantic theory, nor that the results will focus solely on semantics. We focus, rather, upon a word's usage. UFA considers large amounts of evidence about usage, through time, as made available by historical corpora, displaying fluctuation in word usage in the form of a graph. The paper provides guidelines for the interpretation of UFA graphs and provides three short case studies applying the technique to (i) the analysis of the word *its* and (ii) two words related to social actors, WHORE and HARLOT. These case studies relate UFA to prior, labour intensive, corpus and historical analyses. They also highlight the novel observations that the technique affords.

Keywords: collocation, time series data, usage change, Early Modern English, non-parametric regression

1. Introduction

This article introduces a methodology, Usage Fluctuation Analysis (UFA), for the diachronic analysis of large historical corpora. The approach could also be used to explore brachychrony (Mair, 1997). UFA looks at the fluctuation of word usage manifested through collocation, i.e. the co-occurrence of words in texts. It rests on two simple assumptions: (i) words co-occurring in the vicinity of other words provide insight into the words' usage (collocation principle) and (ii) the change in the pattern of co-occurrence of words over time can identify points where their usage changes. The technique itself presupposes neither a specific commitment to any one semantic theory, nor that the results will focus solely on

https://doi.org/10.1075/ijcl.18096.mce
International Journal of Corpus Linguistics 24:4 (2019), pp. 413–444. issn 1384-6655 | e-issn 1569-9811
This work is licensed under a Creative Commons Attribution 4.0 license.

semantics: we focus upon words' usage, not simply their meaning. While usage principally relates to observable linguistic reality, the meaning of words is a more abstract construct with direct implications for semantic theory. Nonetheless, the technique described in this article has some obvious common ground with what claim to be more specifically semantic approaches such as distributional semantics (Harris, 1954; Firth, 1957) and vector space models of lexical meaning (e.g. Clark, 2015). Yet the main purpose of our technique is analytical, i.e. it describes large amounts of evidence about word usage, in different contexts, that are available in historical corpora.

The paper is structured as follows. Section 2 addresses the issue of diachronic change in word usage and meaning. Section 3 describes UFA and provides guidelines for the interpretation of the results of the analysis it provides. Section 4 then offers three short case studies showing applications of the technique in Early-Modern English (EModE). It is first applied to the function word *its*, then to the analysis of social actors through the words WHORE and HARLOT. These studies investigate the value of the technique by comparing UFA to existing corpus-based historical analyses.

2. Language change over time

Word usage analysed through collocation reflects different types of relationships in language, such as semantic, grammatical and discourse-related ones. In historical corpora, collocation enables us to explore the change over time that affects such relationships and their interaction. For example, if we compare the collocations of the word *web* in BNC1994, a 100-million-word corpus that represents British English usage more than twenty years ago (BNC XML edition, 2007; see Table 1) with current British English usage of the same word as captured in the Spoken BNC2014 (Love et al., 2017; Table 2), an 11.5-million-word corpus of British English sampled around the year 2014. In both tables, the collocation statistics are reported using Collocation Parameter Notation (Brezina et al., 2015). 3a-MI(3), L3-R3, C5-NC5 indicates that the MI-score with the cut-off point three was used as the association measure; the collocates were identified in the span of three words to the left (L3) and three words to the right (R3) of the node and the frequency threshold was five for both the collocate (C5) and the collocation (NC5).

The comparison highlights a number of interesting differences. While the literal meaning demonstrated by the collocate *spider* was most prominent in 1994, this association, while still present, dropped to the no. 3 position in the 2014 data. On the other hand, the most prominent current association is connected with the

metaphorical web of the internet as demonstrated by collocates *web* (self-collocation in expressions such as *web design and web development*), *designer*, *development* and *design*. In 1994, another metaphorical meaning was strongly present and demonstrated through collocates such as *caught*, *centre*, *part* and *into*. This was the meaning of *web* as a complicated abstract pattern, as in the following example: *Civil society is at the centre of this web of inter-relationships* [BNC1994, FAW].

Table 1. Top 5 collocations of *web* in BNC1994, 3a-MI(3), L3-R3, C5-NC5

Rank	Collocate	Frequency	MI score
1	spider	6	10.785
2	caught	9	7.75
3	centre	11	6.562
4	part	5	4.2
5	into	11	3.777

Table 2. Top 5 collocations of *web* in spoken BNC2014, 3a-MI(3), L3-R3, C5-NC5

Rank	Collocate	Frequency	MI score
1	web	12	10.785
2	designer	5	9.855
3	spider	9	9.757
4	development	5	9.102
5	design	7	8.703

This simple comparison highlights a few important features of word usage and corpus evidence. First, there is a certain degree of fluctuation of word use over time with new meanings of words emerging such as the internet-related meaning of the word *web* in the Spoken BNC2014. Second, over time meanings of the same word might still be present and overtly demonstrated (*spider's web*), yet others may be latent and not clearly visible in the data (*web* as an abstract pattern in Spoken BNC2014). Third, the amount and quality of corpus evidence varies over time. As Nevalainen (1999: 499) reminds us, in historical corpus enquiries we often need to make the "best use of 'bad' data", i.e. data that is not fully balanced and comparable but that is the only data that is available for a given historical period. This point goes back to Labov's (1994) discussion of the bad data problem in historical linguistics. For example, Tables 1 and 2 are based on 100- and 10-million-word corpora respectively, with the former corpus (BNC1994) representing both speech and writing, while the latter corpus (Spoken BNC2014) represents only informal speech. This comparison was undertaken because there were insufficient examples of *web* in the spoken section of the BNC 1994 alone

to permit this comparison. Hence, some usage variation in diachronic analyses of this sort may be explained by other factors, not simply language change.

Previous work using large diachronic corpora has shown the potential for investigating distributional evidence about words over time (for an overview of prediction-based word embedding models see Kutuzov et al., 2018). For example, Eger & Mehler (2016) used large diachronic corpora for three languages, American English (COHA), German (Süddeutsche Zeitung 1994–2003) and Latin (Patrologia Latina), to visualise what they claim is semantic change by utilizing dynamic graph models. Hilpert & Perek (2015) also employ COHA to investigate the "many a NOUN" construction in American English (1810–2000). They argue that their motion charts, displaying a two-dimensional vector space, reveal an overall decline in noun types in the construction (across different semantic categories) during the course of the twentieth century.

Yet even if we set issues with 'bad data' to one side, previous research has also shown that not all apparent changes in usage herald a change in core meaning; it is possible that as the needs of a speech community vary over time some usages will be more or less preferred. Baker et al. (2017) demonstrate how, over time, the way in which the Ottomans were talked about in the seventeenth century varied. Using the framework of lexical priming theory (Hoey, 2005), Baker et al. (2017: 54) argue that

> primings may be productive when discourse, and world events, give cause for a priming to be used. When those events do not give cause for these productive primings to be expressed, the primings are still present and may act as receptive primings.

Another example shows this clearly – a study of Ireland and the Irish in the UK Hansard corpus found major discontinuities of usage (Baker et al., 2017). One of those, in the 1840s, saw Ireland associate itself in a relatively transient way with a new usage, relating to famine, which was eventually dropped. Because of the scale and importance of the Irish potato famine at this time, *ireland* became strongly associated with the issue but this was not an enduring change in the usage of the word, it was a transient change in usage driven by external events, visible in discourse, which, as events changed, was discarded.

So, while word meaning is one feature of the dynamic nature of a word over time, so is the usage of that word – the word may attract new meaning, lose meaning, be subject to grammatical change, or have meanings shift in and out of productive use depending on circumstance, for example. It may also be subject to other forms of change, e.g. change relating to discourse and changes originating in shifts in pragmatics, which will be visible through collocation.

The process of reusing existing lexical stock represents a challenge for automatic language analysis, which usually performs well when counting forms, but deals with the recognition of the nuances of word usage with only limited success. The accessibility of large datasets of historical language such as the EEBO-TCP corpus (http://www.textcreationpartnership.org/tcp-eebo/), which consists of over a billion running words of English from the mid-fifteenth to early eighteenth centuries, brings this issue to the fore. We can reasonably expect that in a rich time series of linguistic data such as EEBO, we will see word usage shift – indeed McEnery & Baker (2017a, 2017b) have demonstrated this for words relating to prostitution and poverty respectively. However, given the scale of the data becoming available for the analysis of historical data, some tool which at least helps guide the analyst to where usage seems to be in flux or stable would be useful.

In previous work on the EEBO corpus we have, as noted, encountered the dynamic usage of words over time. Our initial attempt to explore this issue came in McEnery & Baker (2017a) in which we looked at how the concept of collocation needed to change to take account of the dimension of time – if collocation is a window onto word meaning and usage (e.g. Brezina et al., 2015; Gablasova et al., 2017), then it follows that it is no more possible to talk about a word having static collocates than it is to talk of a word having static usage. While at any point in time both may appear fixed, through time they are prone to change. This spawns a host of questions of interest to linguists such as when and why did the change occur, how often did the word change its usage, how enduring were the usages attached to the word, what was the nature of the change and how stable or otherwise was the usage of the word over time? Our initial approach (see McEnery & Baker, 2017a: 28ff) to exploring these questions, working with the EEBO corpus, was to split the century into manageable chunks and to use those chunks in order to explore four basic forms of collocate. The chunks we decided on were decades, purely for pragmatic reasons – it gave us sufficient data in each sub-section of the corpus to allow for the exploration of the words we were interested in and it made a close reading analysis practicable, i.e. for each word we wanted to explore in the seventeenth century we would need to analyse it ten times over. The separation of collocates into different types was more principled, as discussed by McEnery & Baker (2017a: 25–28). We were interested in change and as such we wanted to see collocates that were relatively consistent through most of the century (consistent collocates), those which were consistent for a period of time in the century but which fell out of use (terminating collocates), collocates which became consistently used during the century (initiating collocates) and collocates which attach themselves only briefly to a word in the century (transient collocates). This basic typology of collocates over time allowed us to begin to explore changes in usage over time and to test their duration. However, we were cautious when doing so.

For example, while our decade long chunks allowed us to explore our data, we also accepted that they may mask some changes. Some may occur within the decade long chunks, some may straddle them. Also, we were mindful that what we were saying about EModE was based on the data we had access to – that in itself could produce distortions which we needed to bear in mind when producing explanations for change, especially in a century where, for example, substantial changes in censorship practices occurred (see McEnery, 2006: Chapter 3 for details). We also had to carefully explore the extent to which what we were seeing was a side effect of the volume of data available (cf. Nevalainen, 1999).

In order to show changes of usage that have to be explained on linguistic and non-linguistic grounds, we developed UFA, which applies mathematical sophistication to historical corpus data to identify shifts in word use over time. That helps the analyst to broadly characterise usage over time and downsample within a large dataset to use expert analysis, employing the tools of corpus linguistics, in order to characterise usage. Accordingly, in the next section we present this technique, which can help guide an analyst to periods in an historical corpus when the usage of a specific word form appears to be altering or is stable. We then present three short case studies designed to demonstrate the efficacy and use of the technique.

3. Usage fluctuation analysis (UFA)

UFA is an attempt to overcome the problems of the existing automatic and manual methods for dealing with shifts in historical discourse and word usage. Instead of offering a fully automated system, the method combines statistical sophistication with manual (qualitative) analysis. The goal of the method is to automatically identify places where usage change occurs which may deserve the attention of an analyst.

UFA has four main components, described in detail in this section: (i) identification of collocates of a word of interest (node) across a period of time, (ii) use of an overlapping sliding window which moves through the time-series data within which collocates are identified, (iii) recursive estimation of the difference between collocates at any two consecutive points in the sliding window and (iv) application of a statistical regression model to the difference estimate. These four steps are carried out automatically, using a series of scripts freely available from *Lancaster Stats Tools online* (http://corpora.lancs.ac.uk/stats/toolbox.php?panel=6&tab=3). The output of UFA is a graph showing the convergence, or divergence, between collocates within sliding windows moving through time. An example graph, demonstrating the usage fluctuation of the word *web* in the seventeenth century, can be seen in Figure 1.

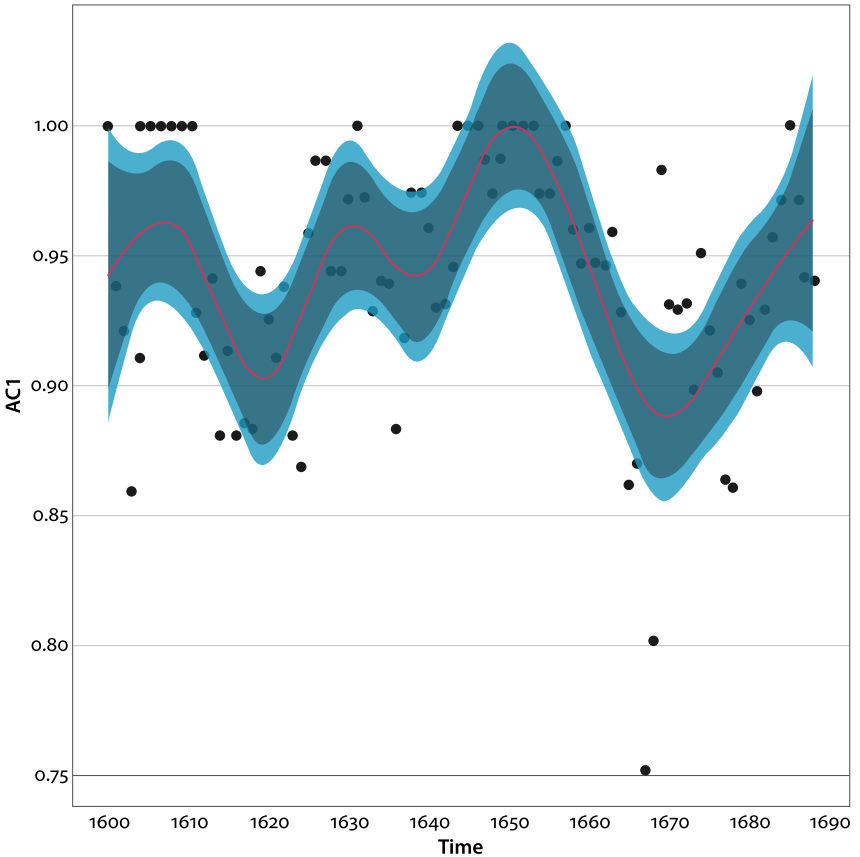


Figure 1. Results of UFA for *web*: 3a-MI(3), L5-R5, $C5_{relative}$-$NC5_{relative}$

In Figure 1 we can see, for example, that the usage of the word *web* underwent a significant change in the 1660s and 1670s, represented by the points lying outside the shaded area in the graph; troughs mark dissimilarity in the set of collocates of the word of interest (*web*) in two consecutive points in the analysis. Different forms of display of similarity and difference are, of course, possible in the graphs. After deliberation, we opted for the convention of similarity being displayed as peaks (UP) and dissimilarity as troughs (DOWN), although a reversed perspective would also be possible.

The tool outputs collocates categorised as consistent, initiating, terminating and transient. The output is shown in Figure 2 below. The dates in the parentheses indicate the period during which a particular item occurs as a collocate with the percentages showing the proportion of the periods during which the collocate occurred out of all considered periods.

Collocates of *web*
CONSISTENT: eye (1600–1687), pin (1600–1690), spider (1600–1690), spiders (1600–1690), weave (1600–1690)
INITIATING:
TERMINATING:
TRANSIENT: away (1624–1681, 27.5%), block (1600–1604, 5.5%), bowels (1683–1690, 8.8%), broad (1672–1681, 5.5%), captain (1616–1622, 7.7%), cloth (1668–1668, 1.1%), curious (1622–1690, 13.2%), d.p. (1689–1690, 2.2%), doctor (1636–1641, 2.2%), exterior (1669–1678, 9.9%), eyes (1625–1659, 24.2%), feather (1669–1671, 3.3%), fine (1666–1666, 1.1%), fly (1622–1624, 3.3%), four (1669–1673, 5.5%), francis (1618–1618, 1.1%), frauncis (1600–1604, 5.5%), george (1600–1604, 5.5%), haste (1605–1614, 11%), her (1630–1630, 1.1%), its (1661–1690, 13.2%), job (1660–1676, 15.4%), like (1616–1690, 30.8%), master (1618–1618, 1.1%), mine (1616–1618, 2.2%), mr (1643–1643, 1.1%), mr. (1678–1678, 1.1%), off (1667–1668, 2.2%), p. (1642–1643, 2.2%), pearl (1630–1681, 15.4%), penelope (1614–1614, 1.1%), pilkington (1618–1618, 1.1%), printed (1635–1637, 3.3%), robert (1600–1604, 5.5%), silver (1600–1604, 5.5%), spin (1618–1690, 23.1%), spindles (1613–1614, 2.2%), spun (1620–1690, 33%), swept (1666–1667, 2.2%), take (1635–1636, 2.2%), takes (1680–1681, 2.2%), thomas (1643–1648, 6.6%), thread (1619–1684, 6.6%), toes (1669–1678, 11%), together (1669–1678, 11%), trust (1667–1667, 1.1%), vessels (1613–1615, 3.3%), weaver (1613–1617, 5.5%), weaving (1634–1637, 4.4%), web (1603–1689, 19.8%), white (1669–1678, 11%), whole (1668–1668, 1.1%), wig (1603–1638, 13.2%), william (1615–1624, 11%), wols (1603–1603, 1.1%), worm (1603–1603, 1.1%), woven (1604–1683, 37.4%)

Figure 2. Collocates categorised by the UFA tool

We can see that there are no initiating or terminating collocates of the word *web* in this period. Most collocates are transient, reflecting specific discourse fluctuations and topics discussed in the source texts. Consistent collocates in this period are connected with the literal meaning of *web* (*spider* and *spiders*), diseases mentioned predominantly in medical texts (*pin and web in mine eye*) as well as the method of production of a web (*weave*). We will return to this example and consider it further in Section 3.5.

Returning to the four components of the UFA procedure, these provide principled answers to four main questions related to the diachronic analysis of collocates:

i. What counts as a relevant collocate?
ii. How does one define a relevant historical period?
iii. How does one compare collocate profiles over time?
iv. How does one identify statistically significant points of change?

The answers to these questions are dealt with in turn in Sections 3.1–3.4. Note that the related question of the nature of the change observed is then for the analyst to determine through close reading, as will be shown in Section 4.

3.1 Identification of collocates

In the broadest sense, a collocate is any word that occurs in the vicinity of the node (word of interest). When operationalising collocation in corpus linguistics, we need to specify (i) what we consider the vicinity of the node to be and (ii) what additional criteria (if any) we use to select a relevant subset of collocates. As regards the first criterion, if collocations are identified using a collocation window rather than syntactic dependencies (cf. Evert, 2008), which is preferable for the analysis of discourse (e.g. Baker, 2004), we first have to define the size of the window, i.e. the span. The collocation span determines how many words to the left and right of the node to consider. Different spans have been used. For example, Sinclair et al. (1970/2004) argue in favour of a 4L 4R (4 words to the left and 4 words to the right) span, based on their experiments and understanding of collocation at the time of writing the OSTI report (1970), though it should be noted that their view of the measurement of collocations is largely contested today (Brezina, 2018; Gablasova et al., 2017). For discourse analysis, a larger span (5L 5R) is usually used (e.g. Baker, 2004). Generally, the smaller the span, the greater the focus of the analysis on the most immediate lexico-grammatical patterns; a larger span captures looser associations.

The second point relates to identifying collocates that are relevant for the study. In corpus research, collocates are ranked based on either frequency of co-occurrence with the node, or, more often, based on an association measure such as the MI-score, Log Dice or Delta P. An association measure is a statistic that highlights certain aspects of the collocational relationship such as the exclusivity or directionality of the relationship between the node and the collocate (Gablasova et al., 2017). Typically, top *n* collocates are considered and explored, so we need to make a principled decision about whether a co-occurring word should be identified as an important and relevant collocate. This binary decision can be made when we introduce a threshold. A threshold specifies that only words co-occurring with e.g. a particular frequency and a particular range of values of the association measure will be considered. Although various heuristic thresholds have been proposed for different association measures (e.g. Hunston, 2002), no principled way has so far been established which clearly defines cut-off points in collocational research.

For the purposes of our analysis, we stipulate a threshold value for the association measure and the collocation frequency (cf. McEnery & Baker, 2017a: 28ff).

However, these threshold values have to be stipulated with some flexibility because the amount of data available for each historical period differs considerably. UFA therefore uses a relative minimum collocation frequency; in addition, the association measure is selected in such a way as to minimise the effect of collocation frequency on the measure itself by selecting a measure that does not correlate with collocation frequency such as the MI score. To establish the relative minimum collocation frequency, we first identify a subcorpus, which provides the smallest amount of evidence for collocation due to the lowest number of occurrences of the node in that particular subcorpus. In other subcorpora with a higher frequency of the node, the minimum frequency threshold is proportionally stricter. For example, if the threshold for the smallest subcorpus is set to be at least three co-occurrences of the collocate and the node, then the requirement for a subcorpus that includes twice the number of nodes would be to include at least twice this number of co-occurrences with the collocate (six in this example). This is an important step in the procedure as we wished the technique to focus on the fluctuation in usage of words through collocation and therefore we wanted to set a consistent criterion of acceptability of a collocate throughout the corpus. Similarly, we also wanted to ensure that any pattern that emerged from our analysis was not simply a reflection of the volume of evidence available for a particular historical period.

3.2 Overlapping sliding window

A crucial consideration in any diachronic study is how to analyse time, which is a continuous variable. In reality, there are no time periods such as years, decades or generations. When we talk about these units, we super-impose our assumed periodisation onto linguistic and social reality. To minimise the effect of large, rigid, discrete time periods, which we had identified as a limitation of our initial technique, we move a sliding window through the data and run collocation analysis for each period inside the sliding window. In practice, we take e.g. ten years as our window and move through the corpus year by year with ten years appearing in the window at each point in time. In principle, given sufficient data, the window could be built on smaller periods of time. A key advantage of the sliding window over our previous decade chunk approach is that the sliding window can capture the cumulative effect of language change better than that. Figure 3 illustrates the sliding window procedure.

Even with the sliding window, certain sampling points (years in this example) are assumed through which the sliding window moves. In Figure 3, the initial and the final stage of the sliding window is indicated by a solid-line frame, while

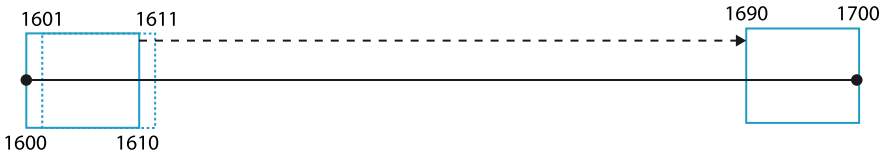


Figure 3. Advantage of overlapping sliding window: it does not presuppose a particular historical periodisation

the movement of the window (dotted line frame) is indicated by the dotted-line arrow.

3.3 Estimation of similarity between collocates

Once the collocates are extracted for each time point inside the sliding window (91 per century when using a ten-year sliding window), a difference between two consecutive points is established. In the literature, different options for computing binary similarity measures have been proposed (Cha, 2007; Choi et al., 2010) including the Jaccard similarity coefficient, Jaccard distance, Sørensen – Dice coefficient and the cosine similarity measure. Choi et al. (2010) review 76 different similarity measures, clustering them according to their performance. In addition, we can also use agreement measures between categorical variables such as raw agreement, Cohen's Kappa or Gwet's AC1 (Brezina, 2018: 87ff). Let us consider the following example. A word X has overall eight collocates at two points in time (t_1 and t_2), some of them are shared, some unique to each data point. At t_1, X collocates with {W_1, W_2, W_3, W_4, W_5, W_6, W_7} and at t_2, X collocates with {W_1, W_2, W_3, W_4, W_5, W_8}. In addition, there are two collocates {W_9, W_{10}}, neither of which co-occurs with X at t_1 or t_2 but which co-occur with X at t_n and therefore are considered as possible collocates. In a collocation matrix, the presence of a collocate is represented by 1, while its absence is represented by 0. A collocation matrix for the example above is displayed in Table 3.

Table 4 shows equations for five selected similarity measures, the values of these measures for the example above (see Table 3) and the basic properties of these measures. These properties focus on whether the measure (i) captures similarity, dissimilarity or agreement; (ii) considers also potential collocates and (iii) considers collocation by chance.

While using a variety of similarity measures would be conceivable, in UFA we prefer Gwet's (2008) AC1, which is an agreement statistic that computes absolute agreement and subtracts from it chance agreement. The advantage of this measure over the other measures listed in Table 4 lies in the fact that it considers the overall agreement between collocates at t_1 and t_2 including potential collocates

Table 3. An example collocation matrix

Collocate	t_1	t_2	
W_1	1	1	
W_2	1	1	
W_3	1	1	a ... present at both t_1 or t_2
W_4	1	1	
W_5	1	1	
W_6	1	0	b ... present at t_1
W_7	1	0	
W_8	0	1	c ... present at t_2
W_9	0	0	d ... absent at both t_1 or t_2
W_{10}	0	0	

that occur at t_n. It also represents the lowest similarity/agreement value (we disregard Jaccard distance for the moment, which is a dissimilarity measure), because it subtracts agreement by chance from the raw agreement value. This is advantageous not only for the theoretical need to control for collocation similarity by chance but also in practical terms because in real collocation matrices the overlap between two points in the sliding window is usually very high. The graphs produced with the AC1 statistic thus most distinctly display the fluctuation in a word's usage. However, there is one direct implication for UFA graphs of using an agreement/similarity statistic rather than a measure of distance such as Jaccard distance: a high value (peak in the graph) represents convergence rather than a point where a shift in usage occurs. Such a point is represented by a low value (trough) (see Section 3.5), though similar distance measures could be computed by subtracting any of the agreement/similarity statistic from 1. In sum, using AC1, we can establish whether the collocates for two consecutive stages of the sliding window agree or disagree and to what extent collocates gradually or rapidly diverge from each other in the course of the time period in question.

3.4 Non-parametric regression model

Finally, a non-parametric regression model (GAM) is applied to the agreement statistic (AC1) data to trace the points where major usage shifts take place. The choice of a statistical model depends on multiple factors. Since mathematical modelling represents a means to an end, it should be guided by the overall aim of the linguistic analysis. Often, the choice of a statistical model is also determined by the procedures preferred in a particular discipline. Realising this instrumen-

Table 4. Selected similarity measures

Measure	Equation	Value for example	Measure of	Considers potential collocates	Considers collocation by chance
Jaccard similarity	$\frac{a}{a+b+c}$	$\frac{5}{5+2+1} = 0.625$	Similarity	NO	NO
Jaccard distance	$1 - \frac{a}{a+b+c}$	$1 - 0.625 = 0.375$	Dissimilarity	NO	NO
Cosine similarity	$\frac{a}{\sqrt{(a+b)\times(a+c)}}$	$\frac{5}{\sqrt{(5+2)\times(5+1)}} = 0.772$	Similarity	NO	NO
Raw agreement	$\frac{a+d}{a+b+c+d}$	$\frac{5+2}{5+2+1+2} = 0.7$	Agreement	YES	NO
Gwet's AC1	$\frac{raw\ agreement - f}{1-f}$ *where* $f = 2 \times \frac{a+b+a+c}{2\times(a+b+c+d)} \times \left(1 - \frac{a+b+a+c}{2\times(a+b+c+d)}\right)$	$\frac{0.7 - 0.455}{1 - 0.455} = 0.45$	Agreement	YES	YES

tal nature of statistics opens up different statistical options for analysing and displaying the similarities and differences of the collocational profile over time. The simplest option is to use moving averages (e.g. Friedman & Stueltze, 1982) – the resulting curve helps smooth extreme fluctuation and describe the main trends in the data. Another option is to use Loess, a local regression model with the traditional method of least-squares. However, the most complex and arguably most flexible option is to use a GAM (Wood, 2017). The advantages of using a GAM include the fact that this model can be extended by adding multiple predictors; it can also deal with missing data and predictions of future lexical development based on past observations. A GAM has also been successfully employed in the 'peaks and troughs' method in newspaper discourse analysis (Gabrielatos et al., 2012). We recognise the need to further validate this methodological choice and perhaps also compare it with the alternative solutions, which might be preferable in certain contexts. Yet, we also believe in the value of introducing GAMs into corpus linguistics – a method that has been successfully used to e.g. model the spread and outbreaks of diseases in epidemiology studies (Kelsall & Diggle, 1998).

While GAMs have many advantages, their use involves specific assumptions, discussion of which is beyond the scope of this article; the reader is referred to specialised reference books such as Wood (2017) for further details. We would note, however, that some of the complexities of using a GAM are common also to the other possibilities mentioned here. For example, the exercise of fitting a smooth curve to fluctuating data is a common feature to all the statistical options outlined above. Similarly, all of them also involve the choice of a smoothing parameter in some form which will directly affect the fit of the curve and the granularity of the analysis.

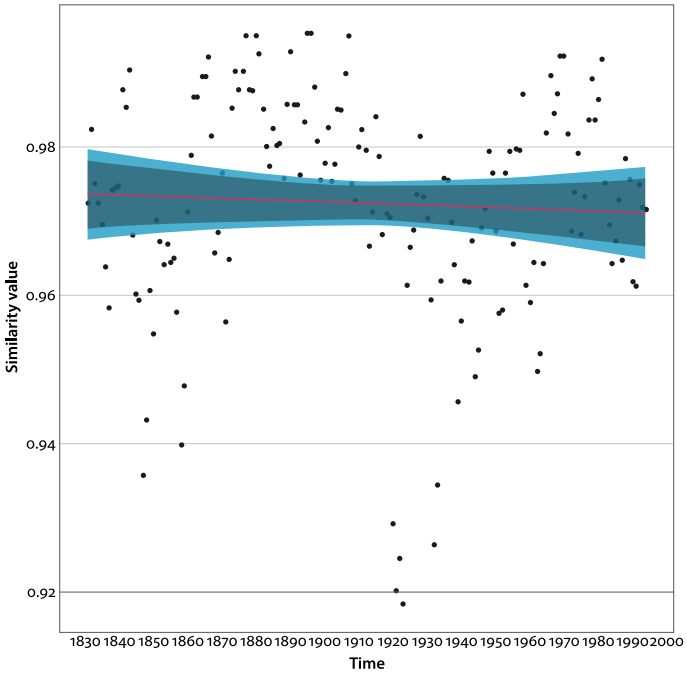
A GAM combines the notion of more widely known generalised linear models (GLM) and additive models, where "the usual linear function of a covariate [is replaced with] an unspecified smooth function" (Hastie, 1990: i). As Wood (2017: xv) points out, there are several practical requirements for using a GAM, namely (i) representation of the smooth function, (ii) controllability of the degree of smoothness and (iii) appropriate selection of the degree of smoothness. The resulting graph is a peaks and troughs graph (cf. Gabrielatos et al., 2012) which shows where the statistically significant points in the development of the usage of a word are, i.e. where the collocates most disagree between two consecutive points (see Figure 1 for an example). We call this trough a point of divergence. These statistically significant points of divergence lie outside the 95% or 99% confidence interval (depending on the alpha level we set for our analysis) for the curve fitted to the data cloud (shaded area in the graph). Figure 4 shows that we have to be careful when selecting an appropriate smoothness parameter to avoid underfitted and overfitted models (Sauerbrei et al., 2006). The statistical model represented

by the curve in the graph needs to be close enough to the data (as in Figure 4c) without reaching to every single data point (Figure 4b), because this approach does not allow generalisation beyond the specific data cloud. On the other hand, the curve needs to reflect the position of the data points better than the line shown in Figure 4a, otherwise it will have little explanatory power. It is also important to consider outliers. Outliers are out-of-the-ordinary values which can compromise the GAM. Gries & Hilpert (2010) propose a method of outlier identification in diachronic data by applying the VNC clustering algorithm, which can highlight clusters with only a single member; these are likely to be outliers. Outliers need to be carefully examined before they can be legitimately removed from the data.

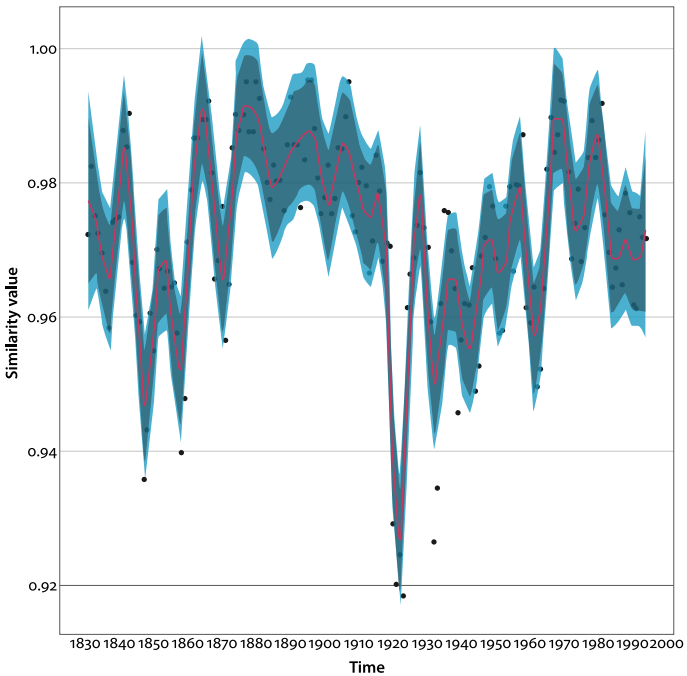
3.5 Reading a UFA graph

Let us consider briefly how to read UFA graphs. We will use the word *web* in the seventeenth century, as displayed in Figure 1, as an example. In all analyses, agreement, disagreement, convergence and divergence are measured by reference to collocation and the similarity between two sets of collocates at two points in an overlapping sliding window (see Section 3.2). It is important to note that the display in the graph is related to the measure of similarity/agreement (see Section 3.3). The specific implications of this display convention are discussed below.

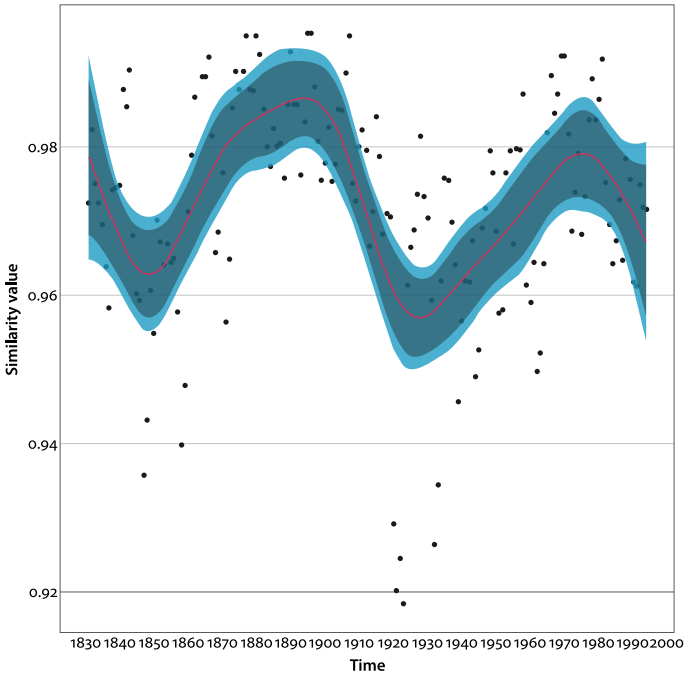
The UFA graph displays the time period on the x axis and the agreement statistic on the y axis. The time period reflects the progress of the sliding window through the data and the series of the pairwise comparisons using the agreement statistics. The dates given on the x-axis indicate the beginning of the first time period considered for the comparison. Thus, for example, the first point in the graph with the mark 1600 in Figure 1 indicates the comparison between the consecutive periods 1600–1609 and 1601–1610. The agreement statistic (AC1 by default) displayed on the y axis of the graph operates on a scale 0–1 with 0 indicating agreement by chance alone and 1 perfect agreement. Typically, points in UFA graphs are closer to 1 than 0, showing overall relative stability of the sets of collocates associated with a node and gradual change in collocates over time. If a point is at the bottom of the y axis (note that the bottom of the y axis does not simply display 0 but shows, rather, the minimum agreement for each given analysis such as 0.75 in Figure 1), there is less agreement between the overlapping windows at that point than any other pairwise comparisons in the particular analysis. If the point reaches the top of the y axis (1), then there is total agreement between the overlapping windows at that point. The higher any point on the graph relative to any other point on the graph, the greater the degree of agreement between collocations at two consecutive points in the sliding window represented by that higher point. Conversely, if



a. Underfitted



b. Overfitted



c. Properly fitted

Figure 4. Different theoretical possibilities of fitting a model to the data

a point on the graph is lower than another point on that graph, then the agreement between the overlapping windows is lower at that point.

Understanding those points allows one to explain how slopes, troughs and peaks work in the graphs. A peak is a local maximum of agreement between the windows being compared in sequence. Similarly, a trough is a local maximum of disagreement produced by the windows being compared in sequence. A peak is preceded by an up slope and followed by a down slope. A trough is preceded by a down slope and followed by an up slope. A down slope marks a period of time during which sliding windows compared in sequence are increasingly divergent. An up slope marks a period of time during which sliding windows compared in sequence are increasingly convergent. One other condition might exist: a plateau. This occurs if the windows being compared are subject to a similar degree of flux, or if the pattern remains consistent across windows, which would be indicated by a plateau where the value on the y axis is 1 consistently.

Note that while we may think of the graphs in these terms, one still needs to move from the graph, through the patterns of collocation and into the data itself to understand the dynamics that are causing the pattern at any point in the graph.

So, the graph guides and controls our investigations – it does not displace or automate them in total. For this reason, UFA in *Lancaster Stats Tools* provides a full, downloadable list of the collocates compared in the analysis showing their presence or absence in each of the periods in question. The relevant collocates can (and should) be further checked in the source corpus using concordancing and other close reading techniques.

4. Case studies

Having described the technique, let us now consider three short case studies which show UFA in operation. To achieve this, we will explore one grammatical word, *its* (Culpeper & Kytö, 2010), and two lexical words, WHORE and HARLOT (McEnery & Baker, 2017a). Our goal in exploring these words is to check for any correspondence or otherwise between the UFA graph and previous studies based on a more labour-intensive approach using fewer sample points.

4.1 UFA analysis of *its*[1]

Culpeper & Kytö (2010) explore *its* building on earlier work by Nevalainen & Raumolin-Brunberg (1994, 2003). Their study, based on sections of the Corpus of English Dialogues (CED; http://www.engelska.uu.se/forskning/engelska-spraket/elektroniska-resurser/a-corpus), the Corpus of Early English Correspondence (CEEC; see https://www.helsinki.fi/en/researchgroups/varieng/corpus-of-early-english-correspondence) and the Helsinki Corpus (see http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/) claimed that *its* undergoes a shift in usage from 1640 onwards (Culpeper & Kytö, 2010: 188), during which it slowly supplants competing choices, such as *of it* and *thereof*, for expressing what Culpeper & Kytö (2010: 185) call the "third person neuter possessive". It is claimed that the choice, in particular in legal discourse, was driven by the selection of *its* as a "less elaborate (more economic) yet also more meaningful (it includes definiteness as well as a signal to supply the referent)" choice (Culpeper & Kyö, 2010: 190).

To explore the evolution of *its* in more detail we had the advantage of more data: the three corpora used in the Culpeper & Kytö (2010) study amount to 5,456,990 words, approximately 0.54% of the size of the EEBO corpus in the seventeenth century. However, we also had the advantage of UFA, which allows us to characterise the development of *its* in functional terms. The results of running UFA on *its* is shown in Figure 5.

1. We would like to thank Jonathan Culpeper for his helpful comments on this section.

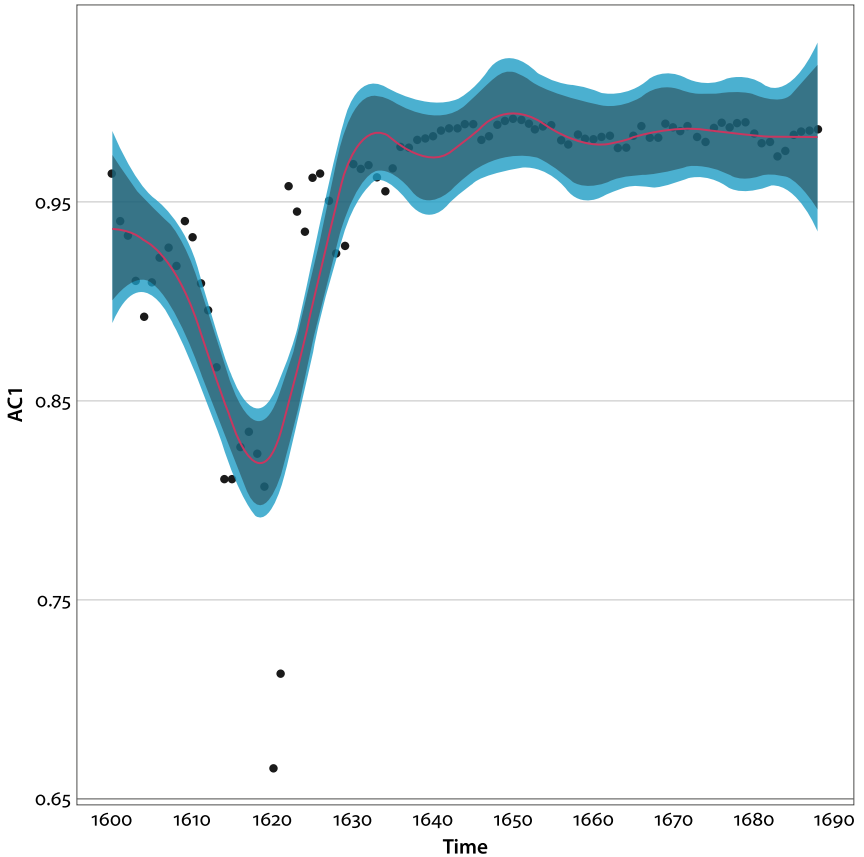


Figure 5. UFA of *its* in the seventeenth century

The graph shows a short period of relative stability for the word at the beginning of the century, with a sharp divergence in the 1610s followed by a period of convergence through to 1630 which leads to an uneven plateau to the end of the century. The period at which *its* undergoes a change in usage is earlier than, but, we will argue, consistent with, the earlier studies of the word. In part, that consistency is based on the difficulty previous studies would have found in observing the change in the 1610s – these studies worked with corpus data in broad blocks (1600–1640, 1640–1680 and 1680–1720) meaning that any fluctuations within a block would be invisible. Yet Figure 5 suggests the change they did observe does occur – so UFA shows that where they could make an observation, UFA may match it, and where they could not make the observation, UFA can add further nuance to the analysis. Let us explore this argument further.

We hypothesised that the UFA graph is showing us a change in the grammatical patterning that *its* will accept. To explore this, we examined the collocates of *its*, looking, in each case, at the modal slot within the window of +/− 5, i.e. that

slot within which the collocate most commonly presents itself, in order to build a picture of the typical pattern of interaction between *its* and its collocates. The patterning is looked at in terms of the part of speech of the word occurring in that slot.[2] Table 5 shows the development of the patterning of the collocates around *its* in the seventeenth century. In the table, only those slots which acquire collocates are shown. The final two decades are omitted from the table as no collocates are added to *its* in these years. Empty cells have no collocates initiating in that position in the decade, otherwise cells show how many collocates initiate with a particular part-of-speech in that decade which have this slot as their modal slot for collocation with *its*. No collocates represented in the table terminate during the period studied, so the total column represents a point of contrast with the 1610s, showing how collocation with *its* has changed across the century.

Table 5. Modal slots for collocations which begin in a particular decade and then persist

Slot	POS	1600s	1610s	1620s	1630s	1640s	1650s	1660s	1670s	Total
−5	APPGE		1							1
−2	NN1							2		2
−2	VVN						1			1
−1	VVI							1		1
−1	VVZ	4	3				1			8
1	JJ	3	3			1		2		9
1	NN1	21	11			7	11	3	1	54
1	NN2	5	1	1		1	2	2		12
2	NN1	6	2							8

What is apparent is that the perturbation in usage shown in the UFA diagram in the 1610s is caused by an intensification of the patterning already associated with *its*. For ease of reference we will call 1600–1620 "Period A". The only innovation in Period A is that *its* begins to collocate with itself ("Again, its nature is known by its colour"; Sala, 1619). By contrast, the changes which continue in what we call "Period B" (which runs principally from the 1640s), hold the degree of convergence in an unsteady plateau as the word acquires collocates fairly steadily through to the 1670s, after which the acquisition of collocates ceases and the plateau stabilizes. It is notable that Period B coincides with and provides evidence for a change in the usage of *its* from the 1640s onwards as has been identified in the literature. This shift is similar to that in Period A in that the existing patterning is strengthened, yet at the same time new patterning attaches to *its* in the −1 and −2 positions.

2. The parts of speech used are those of the CLAWS tagger, see http://ucrel.lancs.ac.uk/claws7tags.html.

Period A establishes two dominant patterns within which *its* appears: with (i) a third person singular verb in the –1 position ("The end of annhiliation, is when a thing loses its present being"; Cuff, 1607) or (ii) a singular common noun ("which also hath its latitude"; Sclater, 1619),[3] a general adjective ("the word in its proper signification"; Humfrey, 1607) or a plural common noun ("let thy face of favour and love spread its beams so over me"; Norden, 1619) in the +1 position. In terms of patterning, while the literature cited discusses *its* as a third person neuter possessive, understanding the patterning of *its* is aided by conceiving of the word as a possessive determiner as outlined in Biber et al. (1999). The second pattern shows very clearly the function of *its* as a possessive determiner which specifies "a noun phrase by relating it to … entities mentioned in the text or given in the speech situation" (Biber et al., 1999: 270–271) as in "Mephibosheth it hath caught a fall, and is lame on its feet" (Everard, 1618), where *its* relates to the proper noun *Mephibosheth*, specifying the ownership of the *feet*. The relationship to a verb in pattern one is also a clear foreshadowing of the use of *its* in Present Day English, where the pattern VERB *its* NOUN is very common. For example, in the Spoken BNC 2014 (Love et al., 2017), 190 (18.02%) of the 1054 examples of *its* in the corpus appear in the pattern VERB *its* NOUN. A third pattern is formed in Period A in the –5 position where *its* collocates with itself. It is worth noting that *its* also self-collocates in the Spoken BNC2014.

Period B strengthens some of the patterns from Period A in the +1 position (adding more collocates which are singular common nouns, plural common nouns and general adjectives). Yet, Period B also extends the patterning around *its* – the –2 position attracts as collocates singular common nouns ("the lawfulness and reasonableness of its use"; Annand, 1661), a past participle ("being divested of its party coloured coat"; Corbet, 1646) and a third person singular verb ("very much contributes to its fertility"; Tavernier, 1677) while the –1 position attracts an infinitive ("awaken the soul to exert its desires"; Hurst, 1678). These results show two processes occurring to the left of *its*. Firstly, it is now being taken as an argument by two phrasal verbs: *divested* of, which accounts for 78 of the 79 examples of the collocate *divested* in this position and *contributes to*, which accounts for 23 of 24 examples of the collocate *contributes* in this position. Secondly, two nouns are associating with *its* (*chyle* and *reasonableness*) in the –2 position. In Period B, in the –2 position, other parts of speech accommodate this possessive determiner – in the case of the phrasal verbs as an argument, in the case of the collocate *reasonableness*, as part of taking a possessive prepositional phrase, through which the noun being modified by the prepositional phrase has the property it possesses specified. In the case of *reasonableness of its*, the noun

3. Which may be dislocated to position +2.

being modified, and the property being ascribed to it, are varied – but the pattern created by the collocation is stable, with 18 of the 22 examples of this pattern fitting the frame *reasonableness* + *of* + *its* + NOUN, where NOUN is the property being ascribed. The collocate *chyle* is linked by a wider range of patterns to *its* – it is linked by verbs (*has*, *lose*, *loses*, *obtains*), a conjunction (*and*) and prepositions (*by*, *for*, *from*, *in*, *through*, *throughout*, *to*, *upon*). While the pattern is more varied, the use of a prepositional phrase to introduce the possessive determiner is still the dominant pattern for *chyle* (14 of 22 examples).

This brief study presents findings which are consistent with previous studies of *its*, yet which also refine and expand them. Some of the refinement is related to the increased volume of data available to this study, but that data is manageable because of UFA – using it, we were swiftly able to downsample and to understand the patterns we saw. In particular, the increased granularity of the sliding window approach allows us to note that the initial period of innovation for *its* occurs in Period A. While the innovation in Period B is attested in the literature, that in Period A is theoretically important – an initial spread of innovation (Period A), followed by a period of consolidation (the 1620s and 1630s) preceding a concerted spurt of diffusion and consolidation (1640–1679) is exactly the behaviour we would expect from changes following the S-curve of linguistic innovation diffusion described by Labov (1994) and other models of innovation as discussed in Nevalaianen (2015). So, as stated, the results of this brief case study are different from, but consistent with, existing work on this word.

4.2 UFA analysis of WHORE

Let us move now to two content words for our next case studies. McEnery & Baker (2017a) analysed words such as WHORE and HARLOT using the decade chunk approach described earlier in this paper. Does UFA replicate, falsify or refine their findings? Replication or refinement would give some indication that the UFA approach is fruitful. Falsification would cast doubt either on the original analysis or on the utility of UFA itself. Figure 6 below gives the UFA graph for WHORE for the seventeenth century for the EEBO data McEnery & Baker (2017a) studied.

In Figure 5, we see that there are two major periods of divergence in usage for WHORE – the first trough indicates that usage for the word begins to enter flux around 1610, with the shift continuing to the mid 1630s until the word veers back towards convergence in the 1640s-1650s. However, the word enters a period of divergence in usage again between 1650 and 1660, with a period of relative convergence following through to the end of the 1670s, before word usage begins to diverge again.

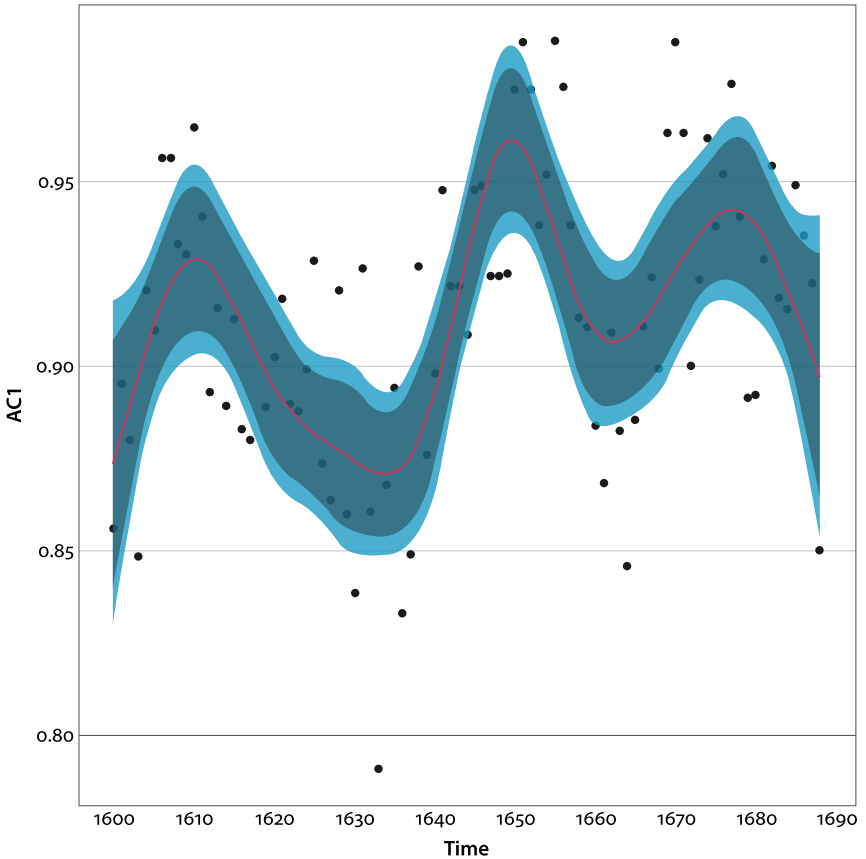


Figure 6. Results of UFA for *WHORE* (lemma): 3a-MI(3), L5-R5, $C10_{relative}$-$NC10_{relative}$

To explore the nature of these fluctuations, we explored the matrix of collocates used to produce the graphs, looking at the collocates generated by each iteration of the sliding window. By contrasting collocates before, during and after the fluctuation, we determined the nature and extent of the change caused by the periods of divergence and convergence. Through concordancing we navigated between specific collocates in a given time range and the examples that produce that pattern of collocation to use close reading in support of our analysis. We then compared our findings back to those produced by McEnery & Baker (2017a: 165–182).

The results here are presented briefly as the UFA led to an analysis which is consistent with that presented by McEnery & Baker (2017a). The role of the word in relation to anti-Catholic sentiment, by reference to the *Whore of Babylon*, is particularly important in explaining the UFA graph in Figure 5. This is the modal usage of WHORE which is sustained throughout the century, as evidenced by consistent collocates such as *antichrist*, *beast*, *cup*, *desolate*, *rome*, *scarlet* and *waters*

(in line with McEnery & Baker, 2017a: 180). The first trough in the graph is caused by a shift to reflect a change in usage associated with what we might call the literal use of the word in the 1630s and 40s, through transient collocates such as *bawds*, *mistress* and *pander*, a result which builds on McEnery & Baker's (2017a: 170–172) discussion of *make* and *pimp*. As these fall away, a peak of convergence occurs in the late 1640s. This begins to descend to a fresh trough as further collocates attach to the *Whore of Babylon* namely *dragon* (which initiates in 1650) and the transient collocate *two-horned* (which occurs in ten consecutive windows starting from 1655). After 1660, a trough of divergence is reached as usages attach to WHORE which are marked and negative (McEnery & Baker, 2017a: 176–181): (i) the word is linked to collocates such as *rogue* (an initiating collocate which attaches to WHORE in 1662) and (ii) the pragmatic affordance of WHORE extends towards being an insult as evidenced through collocates such as *son* (which initiates in 1665). These results align well with McEnery & Baker's (2017a: 172–179) analysis. The UFA shifts across time as different usages of the word attract and shed collocates, the pragmatic affordance of the word extends and the negative discourse prosody of the word intensifies, as discussed by McEnery & Baker (2017a).

4.3 UFA analysis of HARLOT

We will now focus on HARLOT because McEnery & Baker (2017a: 189–193) do not expressly discuss usage change for this word, though they do show the pattern of shifting collocates, at the level of decade, for it. This pattern aligns well with the graph here (see Figure 7) – they show that the 1620s have far fewer collocates (four) than either the preceding (11) or the following decade (12) (aligning well with the first trough in the graph). Similarly, the 1650s represent a peak in the number of collocates attached to the word relative to the preceding and following decade (aligning well with a peak of convergence in the graph). Finally, the 1690s have far fewer collocates for the word (8) than the preceding decade (12) (aligning well with the trough developing at the end of the graph). The UFA graph draws attention to these patterns and demands an answer – it also shows, which the by-decade-approach cannot, that the fulcrum of the change of usage occurs within, not between, the decades.

So what drivers are behind this changing pattern of usage? To explore this, we returned to the procedure outlined – navigating between the collocations that produced this pattern and a close reading of examples. The first thing to be explained is the upwards slope at the start of the graph, indicative of a usage becoming convergent across time. This is apparent in the collocation matrix which is characterized in the analysis which follows.

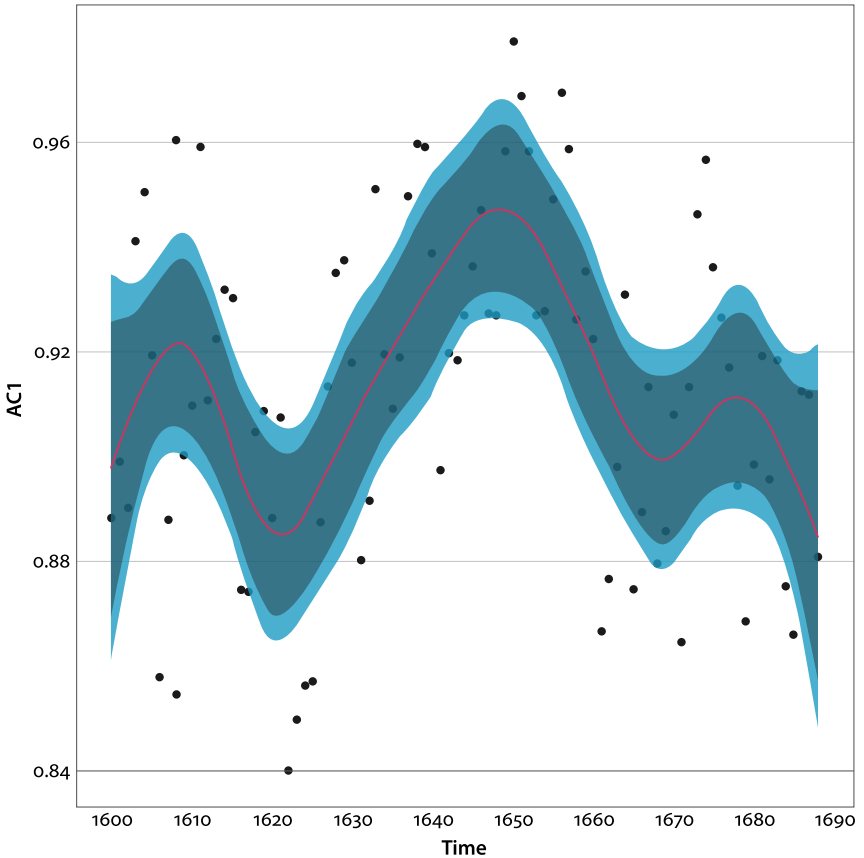


Figure 7. Results of MFA for *HARLOT* (lemma): 3a-MI(3), L5-R5, $C10_{relative}$-$NC10_{relative}$

The first three iterations of the sliding window, which takes us from 1600–1609 to 1602–1611, show a high degree of convergence. There are 20, 23 and 23 collocates respectively in each of these windows. The degree of convergence between these windows increases over time – 16 of the collocates are common to each window, while a further five are common to the second two windows. In contrast, only one collocate is common to the first two windows and similarly only one collocate is shared by the first and last window but not the second window. As the slope suggests, the usage becomes more convergent over time in this part of the matrix. If we focus on the collocates shared by the last two windows, but not the first, we find the words *bawd*, *member*, *notorious*, *repenting* and *wife*. Do these words indicate a change? We can set aside *member* as it is simply the singular form of *members*, which is a collocate shared across all three windows and there seems to be no distinction in meaning between *member* and *members* in the data. The collocate *bawd* clearly links HARLOT to transactional sex, yet in the long run this link is

fleeting – this collocate occurs but twice in the collocation matrix, meaning that the loss of this collocate contributes in part to the second trough in the graph. The collocate *notorious* adds to the negative discourse prosody of HARLOT as established by other collocates in these windows, e.g. *impudent* and *filthy*. Yet it also adds a new dimension of meaning as a notorious harlot is not merely disapproved of but widely known. The collocate *repenting* introduces a discourse of sin and confession to the word – this collocate initiates and sustains up to the 1610–19 sliding window. The detachment and reattachment of repenting to HARLOT occurs for sustained periods throughout the century, contributing to some of the turbulence in the word's usage. Interestingly, the only other collocate that attaches to HARLOT which also clearly links to sin and repentance is the word *sinners*, yet that is only attested in four sliding windows; *repenting* is clearly the main bearer of this discourse, occurring in 41 of the sliding windows in the analysis. Finally, what of the collocate *wife*? In the period covered by the second two windows, *wife* collocates with HARLOT nine times, with all bar one example linking the wife to harlotry as in "thou haste a harlot to thy wife" (Pollock & Holland, 1603). The social drivers for the changes in the first slope are not directly of concern here – what is of concern is that the UFA graph implied a change and, when the collocation matrix was examined and used to downsample to the corpus of texts, a shift in usage was apparent.

To further test the utility of the graph, let us examine another slope, that running from 1650–1662. We isolated the relevant section of the collocation matrix to characterise this shift. For a slope, given that the technique proceeds by pairwise comparison, the relevant windows to start and end an investigation are those which immediately proceed and follow the start and end of the slope. So, for this window, which starts with the window 1650–59 and ends with 1662–1671, the windows from 1649–1658 to 1663–1672 were examined.

This reveals a picture in which the word slowly gains collocates. The word has between 22 (in the first window) and 29 collocates (in the penultimate window) in any given window in this period. In total, the period produces 38 distinct collocates for HARLOT, 19 of which are consistent in this period (*an*, *become*, *city*, *faithful*, *israel*, *joined*, *lovers*, *members*, *mother*, *play*, *played*, *prov.*, *publicans*, *rahab*, *woman*, *abominations*, *forbid*, *judah* and *wife*). In line with the findings of McEnery & Baker (2017a: 161–163), it is obvious even from a cursory investigation of these collocates through concordancing, that the collocates are drawn predominantly from a semantic field of religion.

Six collocates terminate in the period – *young* (first ten windows only), *painted* (in nine of the first 11 windows only), *member* (first six windows only), *plaid* (first four windows only), *hire* (four windows from 1650–1659 to 1654–1663) and *shalt* (first two windows only). Of the remaining 13 collocates, three (*abom-*

inable, *beautiful* and *magdalen*) are transient, occurring in one window only. The remaining ten, however, are present in nine (*roman*, *whore*), eight (*babylon*), five (*haste*), four (*attire*), three (*playing*) and two (*bosom*, *hosea*, *kiss*, *went*) windows. They distribute to the right of the node word. This gives an overall pattern to the activity in the slope. Against a background of a set of stable collocates, some collocates are largely dropped in the first half of the range examined (*hire*, *member*, *painted*, *plaid*, *shalt*, *young*) while another set attach to the word in broadly the second half of the range examined (*abominable*, *attire*, *babylon*, *beautiful*, *bosom*, *haste*, *hosea*, *kiss*, *magdalen*, *playing*, *roman*, *went*, *whore*).

Three of the collocates, *plaid* (which is lost), *played* (which is consistent) and *playing* (which is gained) can be quickly dismissed. They relate to the same lemma and while one might be interested in why one form of the lemma is coming to prominence, the loss of *plaid* is almost certainly a question of changes in orthography rather than meaning *per se*. The relative frequency of *plaid* peaks in the corpus in 1600–1609 at 5.31 examples per million words, generally declining thereafter. By contrast, *played* is always more frequent than that in each decade of the century. Similarly, a single bundle seems to explain *haste*, which typically occurs in the expression "thou haste played the harlot" (36 of the 61 examples of *haste* and HARLOT collocating occur in this phrase).

The key question to ask at this point of the remaining collocates is how systematic and meaningful is the change? Those collocates which are lost do not seem to relate very clearly to those which are gained or maintained, so some degree of change seems to be occurring. This sense increases when those which are gained are compared to those which are maintained – again they generally do not seem to be replicating meaning, they are extending it. Looking at the collocates added in the second half of the range, the nature of the change becomes apparent – HARLOT is being used as a near synonym of WHORE and is often being used to refer to the Whore of Babylon through the collocates *abominable* ("drink the wine of the fornication of that abominable HARLOT"; Savage, 1663), *babylon* ("Cities addicted to Idols, whose Queen is great Babylon Mother of Harlots"; Vilvain, 1654), *bosom* ("And this spirit never can be thus cleansed and [...] enters into the painted bed and bosom of the harlot", Penington, 1659), *roman* ("this Roman Harlot hath made all Nations drunk with the wine of her fornication"; William, 1659) and *whore* (which collocates with HARLOT 37 times in 34 texts and refers to the Whore of Babylon in 34 of the examples). Other collocates index general anti-Catholicism (*attire* "Why should the Spouse of Christ be arrayed in the Attire of an Harlot?"; Latimer, 1661), the Bible (*hosea* is a reference to the Book of Hosea while 31 examples of *went* collocating with HARLOT relate to biblical story in which someone went to a harlot, the other two examples relate to the Whore of Babylon) and temptation (*beautiful* "beautiful Harlots, who after they have had

their lust by men, do many times devour and make them away"; Topsell, 1658, and *kiss* "There is a kiss of lusts and temptation and that is the Harlots kiss", Secker, 1660) and references to Biblical harlots (*magdalen* "remembering Mary Magdalen that penitent harlot"; Trapp, 1649).

Overall, a clear picture emerges – the word HARLOT becomes more closely related with religious usage, and in particular anti-Catholicism, in this period. While it is tempting to explore the motivations and nature of this change further here that is not the purpose of this article. What we have shown here is that the slope in the graph is indicative of change and it provides a useful guide to the analyst who wants to look at change in usage and begin to understand how the change of usage of a word, as evidenced through collocation, allows them to begin to account, in linguistic and historical terms, for changes in word usage over time. In this case, we also show with HARLOT that the increased number of sample points which the sliding window technique provides permits a fuller exploration of usage change than the discrete decade-only analysis of McEnery & Baker (2017a) allowed. While this approach did not invalidate any of their findings, the increased granularity of the findings did permit a more complete picture of the usage of the word to emerge.

5. Conclusion

This article has discussed the methodological principles, application and interpretation of a new corpus analytical technique, UFA. UFA can process large amounts of historical data, focussing on similarities and differences in collocation patterns of different words. It produces summary graphs and collocation tables including automatic categorisations of collocates (consistent, initiating, terminating and transient) according to their occurrence in the period analysed.

The case studies, presented as proofs of concept, show that both function (*its*) and content (WHORE, HARLOT) words are subject to linguistic and social processes that are reflected in the usage fluctuation in corpora. UFA can replicate previous labour-intensive studies reliably while adding granularity. The analysis of HARLOT shows that for words which present relatively sparse data, the sliding window approach may reveal patterns that were not evident in previous studies of the word using static, non-overlapping, windows.

The case studies are illustrative of, but clearly do not exhaust, the uses to which UFA can be put. While the demonstration of the technique showed applications with one specific diachronic corpus (EEBO), the technique itself can easily be applied to any diachronic data that provides enough evidence to explore collocation across time. The work could thus be used to inform, for instance,

ongoing updates of major lexicographic projects such as the *The Historical Thesaurus of English* (https://ht.ac.uk) and the *Encyclopaedia of Shakespeare's Language* (http://wp.lancs.ac.uk/shakespearelang/). As a technique which links an established technique in corpus linguistics (collocation) to a new approach to measuring usage change via a simple interface and easy to understand graphical representation, it is our hope that UFA will provide a gateway to the meaningful exploration of usage change in the wide range of large scale diachronic corpora that are becoming available in a range of languages.

Acknowledgements

The authors gratefully acknowledge the support for the work presented here from the Newby Trust and the Economic and Social Research Council (grant reference ES/K002155/1).

References

Primary sources[4]

Annand, W. (1661). *Panem quotidianum, or, A short discourse tending to prove the legality, decency, and expediency of set forms of prayer in the churches of Christ with a particular defence of the book of common prayer of the Church of England*. London.

Corbet, J. (1646). [no title]. London.

Cuff, H. (1607). *The differences of the ages of mans life together with the originall causes, progresse, and end thereof*. London.

Everard, J. (1618). *The arriereban a sermon preached to the company of the military yarde, at St. Andrewes Church in Holborne at St. Iames his day last*. London.

Humfrey, R. (1607). *The conflict of Job*. London.

Hurst, H. (1678). *The revival of grace in the vigour and fragrancy of it by a due application of the blood of Christ to the root thereof*. London.

Latimer, H. (1661). *The preaching bishop reproving unpreaching prelates*. London.

Norden, J. (1619). *An eye to heaven in earth*. London.

Penington, I. (1659). [no title]. London.

Pollock, R., & Holland, H. (1603). *Lectures upon the Epistle of Paul to the Colossians*. London.

Sala, A. (1619). *Opiologia: or, A treatise concerning the nature, properties, true preparation and safe use and administration of opium*. London.

Savage, H. (1663). *The dew of Hermon which fell upon the hill of Sion, or, An answer to a book entituled, Sions groans for her distressed, &c.* London.

Sclater, W. (1619). *Exposition with notes upon the first Epistle to the Thessalonians*. London.

Secker, W. (1660). [no title]. London.

4. Some titles have been shortened due to space limitations.

Tavernier, J. (1677). *The six voyages of John Baptista Tavernier, Baron of Aubonne*. London.
Topsell, E. (1658). *The history of four-footed beasts and serpents describing at large their true and lively figure, their several names, conditions, kinds, virtues…* London.
Trapp, J. (1649). [no title]. London.
Vilvain, R. (1654). [no title]. London.
William, J. (1659). [no title]. London.

Secondary sources

Baker, H., Brezina, V., & McEnery, T. (2017). Ireland in parliamentary debates. In T. Säily, A. Nurmi, M. Palander-Collin & A. Auer (Eds.). *Exploring Future Paths for Historical Sociolinguistics* (pp. 83–107). Amsterdam/Philadelphia, PA: John Benjamins. https://doi.org/10.1075/ahs.7.04bak

Baker, H., McEnery, T., & Hardie, A. (2017). A corpus-based investigation into English representations of Turks and Ottomans in the early modern period. In M. Pace-Sigge & K. Patterson (Eds.), *Lexical Priming: Applications and Advances* (pp. 42–66). Amsterdam/Philadelphia, PA: John Benjamins. https://doi.org/10.1075/scl.79.02bak

Baker, P. (2004). Querying keywords: Questions of difference, frequency and sense in keywords analysis. *Journal of English Linguistics*, *32*(4), 346–359. https://doi.org/10.1177/0075424204269894

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

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, *20*(2), 139–173. https://doi.org/10.1075/ijcl.20.2.01bre

The British National Corpus, version 3 (BNC XML Edition). (2007). *Distributed by Bodleian Libraries*, University of Oxford, on behalf of the BNC Consortium. Available from http://www.natcorp.ox.ac.uk/ (last accessed August 2019).

Cha, S. H. (2007). Comprehensive survey on distance similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, *1*(4), 300–307.

Choi, S. S., Cha, S. H., & Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, *8*(1), 43–48.

Clark, S. (2015). Vector space models of lexical meaning. In Lappin, S., & Fox, C. (Eds.), *The Handbook of Contemporary Semantic Theory* (pp. 493–522). John Wiley & Sons: Oxford. https://doi.org/10.1002/9781118882139.ch16

Culpeper, J. & Kytö, M. (2010). *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge: Cambridge University Press.

Eger, S., & Mehler, A. (2016). On the linearity of semantic change: Investigating meaning variation via dynamic graph models. *arXiv preprint arXiv:1704.02497*.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (pp. 1212–1248). Berlin: de Gruyter.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis* (pp. 1–32). Oxford: Blackwell.

Friedman, J. H., & Stuetzle, W. (1982). Smoothing of Scatterplots. *Technical Report ORIONo06*. Stanford, CA: Stanford University, Dept. of Statistics. https://doi.org/10.21236/ADA119814

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, *67*(S1), 155–179. https://doi.org/10.1111/lang.12225

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. https://doi.org/10.1075/ijcl.17.2.01gab

Gries, S. T., & Hilpert, M. (2010). Modeling diachronic change in the third person singular: A multifactorial, verb-and author-specific exploratory approach. *English Language & Linguistics*, *14*(3), 293–320. https://doi.org/10.1017/S1360674310000092

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*, 29–48. https://doi.org/10.1348/000711006X126600

Harris, Z. (1954). Distributional structure. *Word*, *10*(23), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hastie, T. J. (1990). *Generalized Additive Models*. Boca Raton, FL: CRC.

Hilpert, M. & Perek, F. (2015). Meaning change in a petri dish: Constructions, semantic vector spaces, and motion charts. *Linguistics Vanguard*, *1*(1), 339–350. https://doi.org/10.1515/lingvan-2015-0013

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. Routledge: London.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524773

Kelsall, J. E., & Diggle, P. J. (1998). Spatial variation in risk of disease: A nonparametric binary regression approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *47*(4), 559–573. https://doi.org/10.1111/1467-9876.00128

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. In E. M. Bender, L. Derczynski & P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1384–1397). Santa Fe, NM: Association for Computational Linguistics. Retrieved from https://aclweb.org /anthology/papers/C/C18/C18-1117/ (last accessed August 2019).

Labov, W. (1994). *Principles of Linguistic Change, Volume 1: Internal Factors*. Oxford: John Wiley & Sons.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014. *International Journal of Corpus Linguistics*, *22*(3), 319–344. https://doi.org/10.1075/ijcl.22.3.02lov

Mair, C. (1997). Corpora and the study of the major varieties of English: Issues and results. In H. Lindquist, S. Klintborg, M. Levin & M. Estling (Eds.). *Papers from MAVEN 97* (pp. 139–158). Växjo: Växjo Universitet.

McEnery, T. (2006). *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. London and New York, NY: Routledge.

McEnery, T., & Baker, H. (2017a). *Corpus Linguistics and 17th-century Prostitution: Computational Linguistics and History*. Bloomsbury: London.

McEnery, T., & Baker, H. S. (2017b). The poor in seventeenth-century England: A corpus based analysis. *Token: A Journal for English Linguistics*, *6*, 51–83.

Nevalainen, T. (1999). Making the best use of 'bad' data: Evidence for sociolinguistic variation in Early Modern English. *Neuphilologische Mitteilungen*, *100*(4), 499–533.

Nevalainen, T. (2015). Descriptive adequacy of the S-curve in diachronic studies of language change. *Studies in Variation, Contacts and Change in English*, *16*. Retrieved from http://www.helsinki.fi/varieng/series/volumes/16/nevalainen/ (last accessed August 2019).

Nevalainen, T. & Raumolin-Brunberg, H. (1994). Its beauty and the beauty of it: The standardization of the third person neuter possessive in Early Modern English. In D. Stein & I. Tienen-Boob van Ostade (Eds.), *Towards a Standard English, 1600–1800* (pp. 171–216), Berlin: de Gruyter.

Nevalainen, T. & Raumolin-Brunberg, H. (2003). *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. Harlow: Pearson Education.

Sauerbrei, W., Meier-Hirmer, C., Benner, A., & Royston, P. (2006). Multivariable regression model building by using fractional polynomials: Description of SAS, STATA and R programs. *Computational Statistics & Data Analysis*, *50*(12), 3464–3485. https://doi.org/10.1016/j.csda.2005.07.015

Sinclair, J., Jones, S., & Daley, R. (2004). *English Collocation Studies: The OSTI Report*. London: Bloomsbury. (Original work published 1970)

Wood, S.N. (2017). *Generalized Additive Models: An Introduction with R*. Boca Raton, FL: Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279

Address for correspondence

Tony McEnery
Department of Linguistics and English Langauge
Lancaster University
Bailrigg
Lancaster, LA1 4YL
UK

Co-author information

Vaclav Brezina
Department of Linguistics and English Language
Lancaster University

Helen Baker
Department of Linguistics and English Language
Lancaster University