Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition

Martin Hilpert
Freiburg Institute for Advanced Studies (FRIAS),
Freiburg, Germany

Stefan Th. Gries
University of California, Santa Barbara, USA

Abstract

The use of corpora that are divided into temporally ordered stages is becoming increasingly wide-spread in historical corpus linguistics. This development is partly due to the fact that more and more resources of this kind are being developed. Since the assessment of frequency changes over multiple periods of time is a relatively recent practice, there are few agreed-upon standards of how such trends should be statistically interpreted. This article addresses the need for a basic analytical toolbox that is specifically tailored to the interpretation of frequency changes in multistage diachronic corpora. We present a number of suggestions for the analysis of data that analysts commonly face in historical studies, but also in the study of language acquisition.

Correspondence:
Martin Hilpert,
Freiburg Institute for
Advanced Studies (FRIAS),
Albertstr. 19,
79104 Freiburg, Germany
E-mail:
martin.hilpert@frias.
uni-freiburg.de

1 Introduction

The use of corpora that are divided into temporally ordered stages, so-called *diachronic corpora*, is becoming increasingly wide-spread in historical corpus linguistics (Lindquist and Mair, 2004; Kohnen, 2006; Lenker and Meurman-Solin, 2007; amongst many others). This development is partly due to the fact that more and more resources of this kind are being developed, especially with regard to English (Beal *et al.*, 2007). The Helsinki Corpus, which can be said to have pioneered the genre, has been substantially expanded, and numerous other corpora now offer comparable sets of texts that represent subsequent periods of time in the development of a language. Similarly, many studies of language acquisition have been relying on resources like the CHILDES corpora (MacWhinney, 2000), which also allow comparisons between sequentially ordered periods of time.

Since the comparison of frequency values over multiple periods of time is a relatively recent practice, there are few agreed-upon standards of how observed frequency changes in diachronic data should be statistically interpreted. This does of course not mean that this subject has gone unexplored. As quantitative analytical methods are central to current sociolinguistics, it comes as no

Literary and Linguistic Computing, Vol. 24, No. 4, 2009.

doi:10.1093/llc/fqn012 Advance Access published on 1 September 2008

surprise that there have been some variationist approaches to the analysis of diachronic corpora (Biber and Burges, 2000; Nevalainen, 2000). Other suggestions for the quantitative analysis of diachronic corpus data can be found in Gries and Hilpert (2008), Hilpert (2006), and Hinneburg *et al.* (2007). More often than not, however, frequencies reported in historical studies merely serve the purpose of illustration and are not subjected to any further statistical analysis. If illustration alone is the ultimate goal, then there is nothing to be said against this practice. However, in this article, we make the case that there are insights to be gained from exploratory statistical techniques, which may reveal phenomena that are not observable through mere eyeballing of frequency data. Often enough, trends are not unidirectional, or not strong enough to be intuitively clear. Trends may also increase or decrease in strength over time. Whether a trend has become significantly more pronounced over time can be a nontrivial question, so that it would be desirable to have a method of assessment that is data driven, rather than based on intuitive judgments. It is the goal of this article to develop methods of this kind and to explain their application with illustrating examples.

The remainder of this article is organized as follows. Section 2 first details some general characteristics of work with diachronic corpora and then describes five basic analytical scenarios, discussing in each case the questions that hinge on observable frequency developments. Section 3 offers for each of these scenarios a statistical procedure to assess and interpret the available information. Section 4 concludes the article and offers an outlook to further research.

2 Analysis Types for Multistage Diachronic Corpora

What are the problems that can be approached through the analysis of diachronic corpora? To answer this question, we need to discuss some particulars of the work with such resources. A diachronic corpus is, first and foremost, a collection of texts that vary along the parameter of time.[1] Along the time axis, the corpus compiler (or the researcher) makes a number of essentially arbitrary boundaries to divide the corpus into successive periods. To take an example, the TIME corpus (http://corpus.byu.edu/time/) is divided into nine periods that represent decades from the 1920s up to the first decade of the 21st century. The first period represents the years 1923–29, the last period represents the years 2000–06, so that not every period covers, strictly speaking, a decade. The TIME corpus holds about 106 million words, divided into different amounts of words for each decade, as shown in Table 1 (these frequencies are based on online searches in December 2007).

The varying sizes for each period necessitate a normalization of observed raw frequencies into measures such as instances per million or per 10,000 words, if a frequency development of a single form, say, the word *internet*, is to be evaluated. If we are interested in the frequency of one form relative to another, such as for instance the frequency of *keep* with a gerund complement (e.g. *keep moving*) relative to the frequencies of other uses of *keep*, the varying corpus sizes need not concern us any further, since the frequency ratios between these forms can be calculated independently of the overall corpus size. The most basic observation that can be made about the frequency development of a given form is whether it became more or less common, or whether it remained relatively stable. Trivial as this may seem, it is not always obvious whether an observed trend constitutes a significant development or an accidental fluctuation in the data. Table 2 illustrates this problem with five examples from the TIME corpus.

2.1 *in*

First, let us consider an element like the preposition *in*, for which we do not hypothesize to find major

Table 1 Words in the sub-periods of the TIME corpus (in million words)

Time period	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s
Words (in million)	7.4	12.3	15	16.2	15.7	12.5	11.1	9.4	6.7

frequency developments. Eyeballing the normalized frequencies seems to confirm that no drastic changes occurred—we see a moderate increase during the 1950s and 1960s, but after that the frequencies approximate their earlier values. However, is there a way to assess more formally whether the development represents common fluctuation or whether it reflects a genuine phenomenon that should not be dismissed that easily? Put more concisely, is there a statistical measure that would tell us if one or more observed frequencies deviate more strongly from the mean than we could reasonably expect?

2.2 *and*

This question can be further illustrated with the conjunction *and*, which seems to undergo a slight increase in frequency over the nine periods of the TIME corpus. In assessing this trend, the analyst faces (at least) two possible scenarios: the null hypothesis would be that we observe common fluctuation; the more interesting research hypothesis would be that this trend actually has a reason. One candidate for such a reason would be the on-going colloquialization of written English (cf. Leech and Smith, 2006; Mair, 2006; Kohnen, 2007). On the colloquialization hypothesis, complex patterns of syntactic subordination gradually give way to paratactic structures, a tendency that would be consonant with an increase of *and* in written corpus data. Since now a theoretical issue is at stake, we want to determine with the greatest possible certainty whether the observed frequency increase could be due to mere chance.

2.3 *whom*

Of further interest are cases in which it is intuitively clear that a noteworthy change has occurred. In such cases, we would like to be able to describe the dynamics of that change in more precise terms. To illustrate, the relative pronoun *whom*, the third element shown in Table 2, has undergone a substantial decrease in text frequency since the 1920s. Ideally, we would like to arrive at a more fine-grained result than just the finding that the change between then and now is statistically significant. From eyeballing the numbers we can derive the hypothesis that *whom* really just underwent a frequency decrease in the first four periods and remained stable after that. This would allow the conclusion that the present-day use of *whom* is confined to a narrow range of contexts, which nonetheless provide a relatively safe ecological niche for it. Is there statistical evidence that would allow us to group different periods of the corpus together and make individual statements about frequency developments in each of these groups?

2.4 *just because . . . doesn't mean*

Another illustration of a dynamic trend is the development of the so-called *just because . . . doesn't mean* construction, which is exemplified by sentences such as *Just because they have McDonald's and Barbie dolls, we shouldn't expect they'll think and act like Americans*. A look at the frequencies shows that the construction was used with an evenly low frequency until the 1950s. After that, there is a steady increase that appears to become greater after 1980.

Table 2 Frequency developments of five expressions in the TIME corpus

	1920s	**1930s**	**1940s**	**1950s**	**1960s**	**1970s**	**1980s**	**1990s**	**2000s**
in n	139,528	215,243	294,086	342,866	346,270	251,004	214,784	175,120	129,908
per 10k words	188.72	174.79	196.15	211.07	221.24	200.52	194.32	185.78	192.54
and n	149,434	185,543	285,890	358,437	379,730	283,265	236,864	213,033	157,476
per 10k words	202.11	150.67	190.69	220.66	242.62	226.29	214.29	226.01	233.40
whom n	2000	2987	2737	2419	2675	1972	1463	1486	989
per 10k words	2.71	2.43	1.83	1.49	1.71	1.58	1.32	1.58	1.47
just because n	4	5	7	7	16	11	14	27	26
per 10k words	0.005	0.004	0.005	0.004	0.010	0.009	0.013	0.029	0.039
keep V-*ing* n	97	251	613	939	782	599	585	727	523
keep other n	2658	5433	9205	10,012	8902	5821	4564	4157	3425
% *keep V-ing*	0.04	0.04	0.06	0.09	0.08	0.09	0.11	0.15	0.13

As with the data for *whom*, we would like to be able to capture significant differences between stages in the development of the construction.

2.5 *keep*

The above examples concern only frequency changes of individual linguistic forms. Some cases merit simultaneous attention to the frequency developments of two or more forms, as such a perspective can shed light on commonly encountered scenarios of grammatical change. An illustrating example from English is the verb *keep*, which has a number of lexical senses relating to the idea of retaining a state (*keep quiet*), object (*keep the money*), or location (*keep off*), but which in conjunction with a gerund complement has come to express the grammatical category of continuative aspect, as in *He keeps telling me about his problems*. From a quantitative perspective, it would be interesting to see whether the ratio of examples in which *keep* takes a gerund complement has increased over time. Table 2 shows that this is indeed the case. Whereas, the form *keep V-ing* only accounts for 4% of all uses of *keep* in the 1920s, its relative frequency compared to other uses of *keep* has more than tripled in the 1990s and 2000s.

3 Analysis Techniques for Multistage Diachronic Corpora

Facing the kind of data described in the previous section, the obvious question is how one is to characterize the development of the words/phrases over time, which raises the following issues:

- Is there one overall trend in the data and how do we find that out?
- If there is any one overall trend, what kind of trend is it: upwards, downwards, or stable?
- Are there several sections or parts in the data that exhibit commonalities that set them apart from other parts of the data and how do we find that out?

In the following sections, we will discuss several different approaches to these questions. We think that these approaches are best followed in a step-by-step fashion, which is why our exposition here will have a manual-like character even if for reasons of space we cannot discuss all approaches with all five expressions. The common denominator of all these approaches is that they are all bottom-up/data-driven in the sense that we try to minimize the effect of subjective preconceptions of individual researchers. Further, the approaches are quantitative in the sense that they are based on statistical methods. While we admit that this entails a certain degree of technicality, (1) fields other than linguistics have long been using much more complex techniques than the ones we outline below, and (2) we believe that the diagnostic value and the degree of objectivity that comes with such approaches makes these approaches worth their while. We will discuss a method to determine whether the data as a whole exhibit a particular trend in Section 3.1; two methods to determine whether there are additional sub-structures in the data (one of these will be based on pairwise differences between successive observed values in Section 3.2; the other will be based on the observed values as such in Section 3.3); and a method for the analysis of interrupted trends in Section 3.4.

Some of the example discussion below is easier to understand on the basis of visualized data, so let us first show how the development over time of the five expressions looks like in simple scatterplots. Consider Fig.1, where the *x*-axis represents the passage of time and the *y*-axis represents the observed frequencies of the expressions we have discussed above. For *in*, *and*, *whom*, and *just because*, the *y*-axis represents tokens per 10,000 words, for *keep V-ing*, the *y*-axis shows its relative frequency compared to other uses of *keep*. The solid lines indicate the developments of the frequencies over time; the dashed lines are nonparametric smoothers summarizing the developmental trends.

3.1 The detection of trends

The first analytical step is to look at the simplest question: is there one overall trend in the data? The simplest conceivable approach to answer this question involves the use of rank-order correlations. Does the sequential order of the different corpus periods correlate with a ranking of some kind?

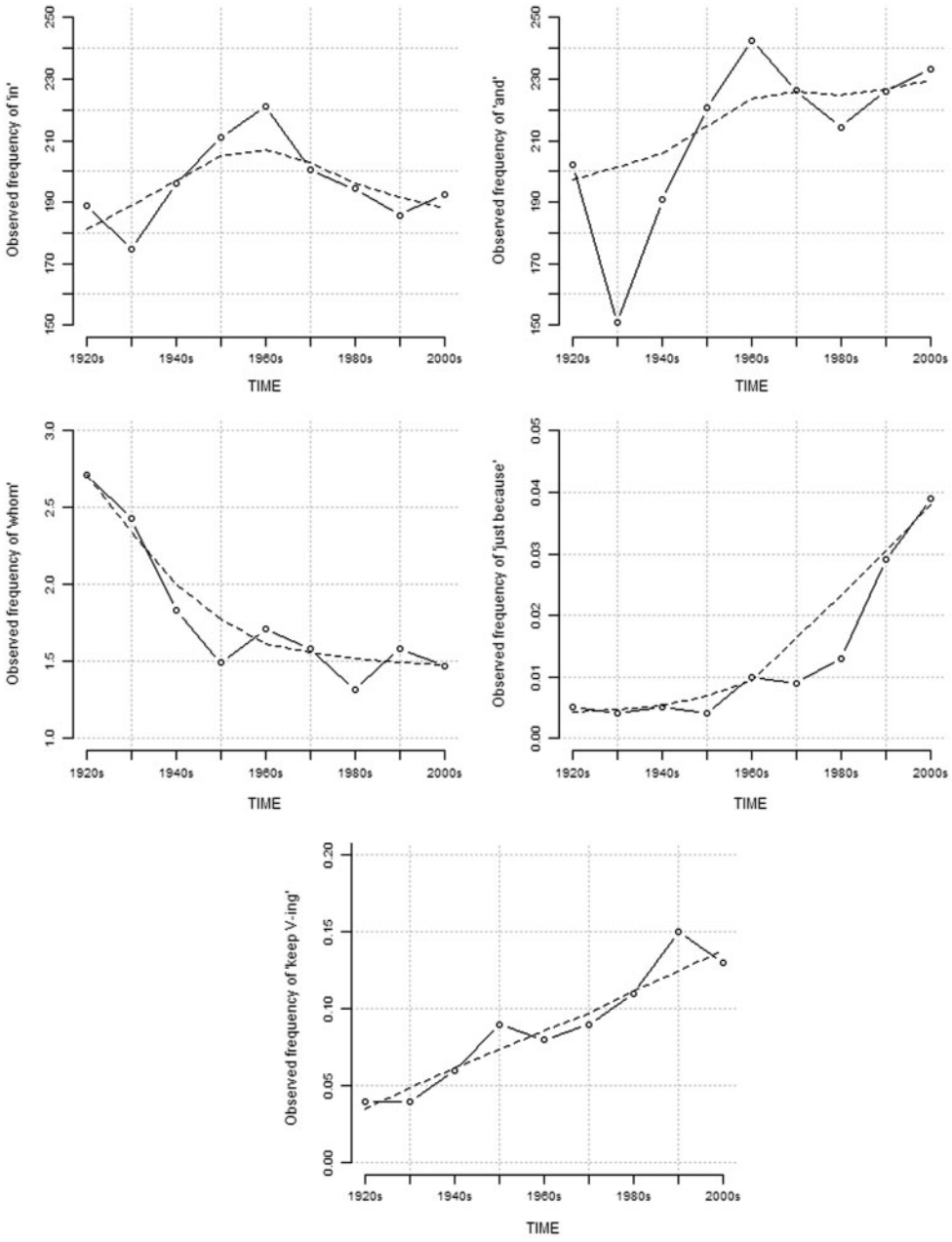

Fig. 1 Scatterplots representing the developments of frequencies

For each of our examples, we have nine data points, their relative frequencies. A perfect upwards trend would mean that each observed frequency at time point x would be higher than that at time point $x-1$, and of course the reverse would hold for a downwards trend. A wide-spread measure for correlations are coefficients such as the Pearson product-moment correlation or Kendall's-τ. Since the former presupposes interval data and is more sensitive to outliers than the latter, we

Table 3 Kendall's τ for the relative frequencies of the five expressions and nine time periods

	in	and	whom	just because	keep V-ing
Kendall's τ	0	0.5	−0.704	0.743	0.857
$P_{\text{two-tailed}}$	1	0.075	0.008	0.005	0.001

will use Kendall's-τ here. Correlating the sequence of corpus sub-periods (1–9) with the relative frequencies from Table 2 above produces correlation coefficients for each example, which are shown in Table 3.[2] A value close to 0 indicates the absence of a trend, values approaching either 1 or −1 indicate that the passage of time correlates perfectly with an increase or decrease in frequency.

The results in Table 3 provide a first indication of what may be relevant. They show, for example, that *in* does not exhibit any trend at all, that *and* exhibits an intermediate trend that is marginally significant, while the other expressions exhibit very significant upward trends (*just because* and *keep* V-*ing*) and downward trends (*whom*). These are of course no earth-shaking results, since the numbers merely confirm to some degree what one would have reasonably inferred from the graph. However, even here the statistical evaluation already goes beyond what eyeballing the graph can do. First, it would be hard to guess from the graph alone that the trend for *and* is only marginally significant. Second, it would be hard to compare the strengths of the trends of *whom* and *just because*—only the numbers tell that the former is slightly weaker than the latter.

3.2 Variability-based neighbor clustering with the actual values

The second step in our 'manual' is to investigate the internal quantitative structure of the data in more detail. Knowing whether there is or is not a trend is good, but far from sufficiently precise. Thus, this step needs to be done regardless of whether there is a trend in the data or not:

- if there is a trend, then the correlation coefficient from above does not reveal what the exact nature of the trend is like: even if Kendall's-τ is significant, the trend may not be linear (which can often already be gleaned from the scatterplot); also, the trend may involve different steps or parts;
- if there is no overall trend, there may still be identifiable structure in the data that is worth exploring.

The second step is variability-based neighbor clustering (VNC), a method that was originally developed for the identification of stages in temporally ordered data in language acquisition (cf. Stoll and Gries, forthcoming) and diachronic linguistics (cf. Gries and Hilpert, 2008), but that can also be used as a more general heuristic to identify structures in different kinds of data. VNC is a hierarchical cluster-analytic approach, but unlike regular clustering methods, it takes into account the temporal ordering of the data. Thus, it groups together data from different time periods on the basis of their similarity, but only merges data points that are immediately adjacent (hence 'neighbor' clustering), effectively preserving the temporal order that characterizes language acquisition data or diachronic historical corpora.

While we cannot discuss all aspects of VNC here (cf. the above references for details), a brief characterization of the iterative algorithm is in order, which we show in Algorithm 1 in pseudocode. Like most iterative algorithms, such procedures unfortunately do not lend themselves well to a characterization in prose.

The output of VNC is a kind of dendrogram familiar from regular clustering approaches, which plots the amalgamation of the nine time periods such that the *y*-axis represents the similarities between different data points and clusters. Let us explore what this algorithm has to offer given two examples from the present data, starting with the example of *just because*. On the first iteration in lines 2 and 3, VNC accesses the first and the second time period (1920: 0.005 and 1930: 0.004) and computes the variation coefficient of these two values (0.1571). It proceeds to do the same for all successive pairs of values, the second and the third, the third and the fourth, etc. always storing the variation coefficients. After that, in line 6 VNC identifies the smallest variation coefficient, which indicates the values that are most similar to each other and thus merit being merged first into

```
Given nine different time periods whose nine values represent relative
frequencies of occurrence of a particular linguistic element and where each
frequency is named according to the (average) year for which this frequency was
observed,

01 repeat
02    for all but the last time period
03       access the rel. freq. from the x-th period
04       access the rel. freq. from the x+1-th period
05       compute a measure of their similarity (their variation coefficient)
06    identify which of the pairs exhibits the largest similarity (the
         smallest variation coefficient)
07    merge the data of the two most similar time periods by
08       computing the weighted mean of the rel. freq. of the time periods
            that are most similar
09       renaming the merged time periods to the weighted mean of the years
            that are merged
10 until all time periods have been merged
```

Algorithm 1 VNC for the TIME data (in pseudocode)

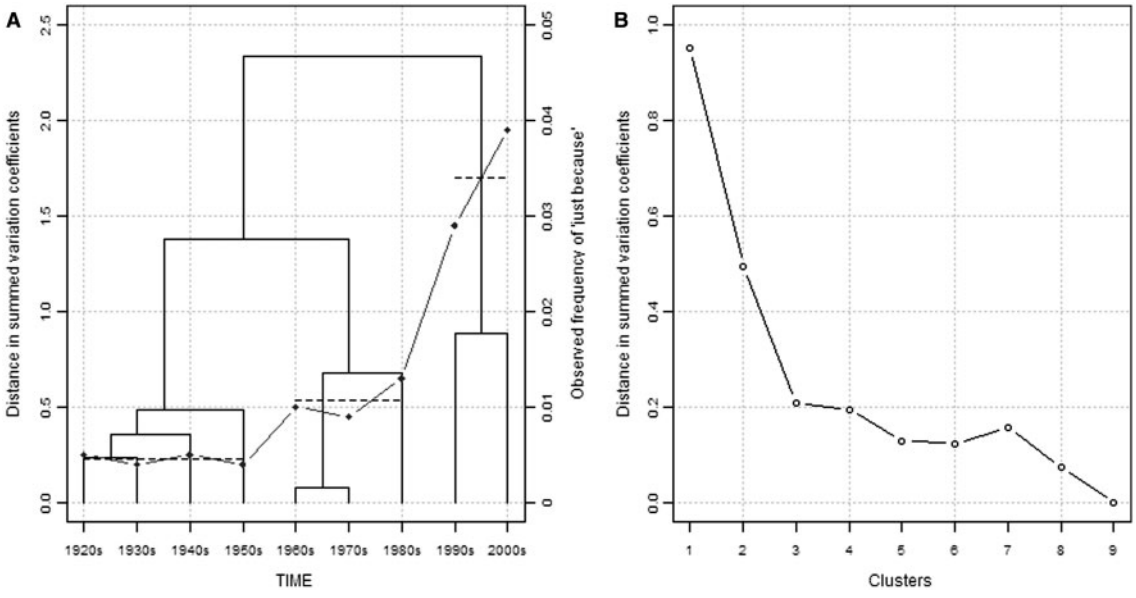

Fig. 2 VNC dendrogram for the TIME data on *just because* (**A**) with overlaid line plots of observed frequencies and mean frequencies per cluster and scree plot (**B**)

one group. In the first iteration, this is the pair of values at 1960 and 1970, whose variation coefficient is 0.0744. In lines 7 to 9, VNC then merges the two data points by computing the means of the data points (0.0095) and the years (1965). Thus, after first iteration, there are not nine data points anymore, but just eight, with the new value of 1965: 0.0095 taking up the place of the former values for 1960 and 1970.

Lines 1 and 10 ensure that this process is repeated until only one data point is left. That is, on the second iteration, VNC again compares all values of successive pairs of time periods to each other, merges the two most similar time periods, etc.[3] The left panel in Fig. 2 plots the amalgamation of the time periods and the distance on the *y*-axis corresponds to the sum of variation coefficients, so we see that 1960 and 1970 are indeed merged first at $y = 0.0744$.

The right panel of Fig. 2 plots the variation coefficients as an analog to scree plots in principal component analyses, where they are used as a guideline to determine how many factors should be included in a model. Here, the plot indicates how many different stages should be assumed within a diachronic development. The plot shows substantial distances between the three largest clusters, i.e. a steep slope between the first three points. After the third cluster, the curve levels off to the right and becomes nearly horizontal. This suggests a division into three separate historical stages, each represented by a cluster. The dendrogram (left panel) reveals what these clusters are. Cluster 1 ranges from the 1920s to the 1950s, cluster 2 ranges from the 1960s to the 1980s, and cluster 3 ranges from the 1990s to 2000s. The dashed horizontal lines indicate the mean frequencies of *just because* that are observed in the data for the three clusters. So, while these data clearly support the finding from Kendall's-τ that there is an upward trend, this approach also provides the additional, more fine-grained information that it is probably most useful to interpret the trend in the data as involving three successive stages with mean frequencies of occurrences of 0.0045, 0.01, and 0.034 for cluster 1, cluster 2, and cluster 3, respectively. In summary, both the dendrogram and the scree plot in Fig. 2 strongly suggest that a characterization of the development of *just because* just on the basis of the overall trend is insufficient since that would fail to note that there are in fact three different stages in the data. Without such an approach, precision and objectivity of this kind are hard to come by. Consider now Fig. 3 for analogous representations of the data for *keep* V-*ing*.

Although *keep V-ing and just because* have similarly high Kendall's-τ values, the respective results of VNC are quite different. The scree plot for *keep V-ing* in Fig. 3 shows that, apart from the overall upward trend that Kendall's-τ already identified, there is hardly any additional structure in the development of *keep* V-*ing*. There is no early point where the scree plot levels off, and it levels off only at so late in the construction of the dendrogram that by that time most clusters consist of only one time period. An analysis of the frequency changes found with *keep* V-*ing* thus need not concern itself with any sub-stages within the overall time frame; reporting the overall trend is sufficient. Still, this finding can

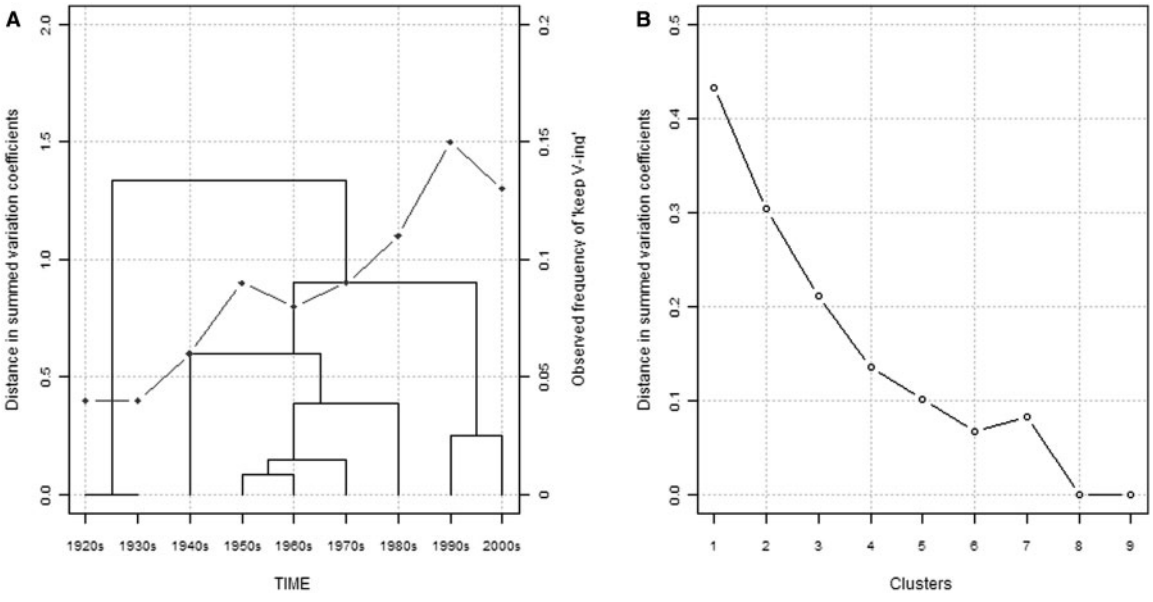

Fig. 3 VNC dendrogram for the TIME data on *keep* V-*ing* (**A**) with overlaid line plots of observed frequencies and scree plot (**B**)

only be obtained through a VNC application that specifically tests for the existence of sub-stages.

3.3 Iterative sequential interval estimation using changes between actual values

In the previous section, we have discussed how VNC could be used to determine whether there are sub-structures in the data. VNC is based on the absolute differences between data points: while it measures the differences between each data point and the immediately following one, it does not take into consideration whether that difference reflects an increase or decrease in frequency. This is necessarily not only a weakness and does result in cluster structures that both match intuitions about the shapes of the curves, but also still provides new information that would be likely to escape the naked eye. In this section, we discuss a second heuristic that also allows the detection of sub-structures in the data, but that does so in a way that is more concerned with the entire development that has led up to a particular point of time.

Just like VNC, this heuristic is an iterative algorithm, called ISIE for *iterative sequential interval estimation*. It repeatedly moves through successively larger sequences of data points to produce continuously updated estimates about how a developmental or diachronic curve will proceed. To put it simply, the method gives us a range of expected values for the next step in the development. If that next step happens to go beyond the expected values, we have detected a change that merits further attention. The procedure is characterized in Algorithm 2. Again, this algorithm may seem rather opaque initially, but will become clear once we discuss an example.

The application of this to the data for *just because* and *keep* V-*ing* results in Fig. 4.

Let us again move through Algorithm 2 on the basis of the data for *just because*. The first seven lines merely gather the pieces of information that all following calculations require: on the first iteration in line 2, ISIE records the difference of the second value minus the first value and generates a vector d, which contains the value of $0.004 - 0.005 = \{-0.001\}$. In line 3, ISIE takes the negative sign from the one observed difference in d and generates a vector s $\{-1\}$. In line 4, ISIE generates the mean of all absolute differences in d, which is of course just 0.001. In line 5, ISIE generates a vector p, which is $\{1\}$ since there is just one data point yet. In line 6, ISIE would add up all values of p, where s is positive, but since there are none sum_+ becomes 0. In line 7, ISIE adds up all values of p, where s is negative so sum_- becomes 1.

With these pieces of information in place, the algorithm can proceed to make upper and lower estimates for the next step in the development.

```
Given 9 different time periods whose nine values represent relative frequencies
of occurrence of a particular linguistic element,

01 for all frequencies but the first and the last
02    generate a vector d with all n pairwise differences between successive
         rel. freqs. from 1 to the current frequency
03    generate a vector s that contains the signs of the n differences (i.e.,
         +1 for increases and -1 for decreases)
04    determine the average of the n absolute differences a, which will
         function as the range of predictions
05    generate a vector p that contains the percentages of the values n to 1,
         out of the cumulative sum of the values 1 to n
06    generate the sum of all percentages in p whose corresponding value in s
         is positive; call that sum+
07    generate the sum of all percentages in p whose corresponding value in s
         is positive; call that sum-
08    draw a line from the observed frequency of the current recording to the
         next recording + sum+ times a
09    draw a line from the observed frequency of the current recording to the
         next recording + sum- times a
10    shade the area between the lines to highlight the predicted range for
         the next frequency
```

Algorithm 2 ISIE for the TIME data (in pseudocode)

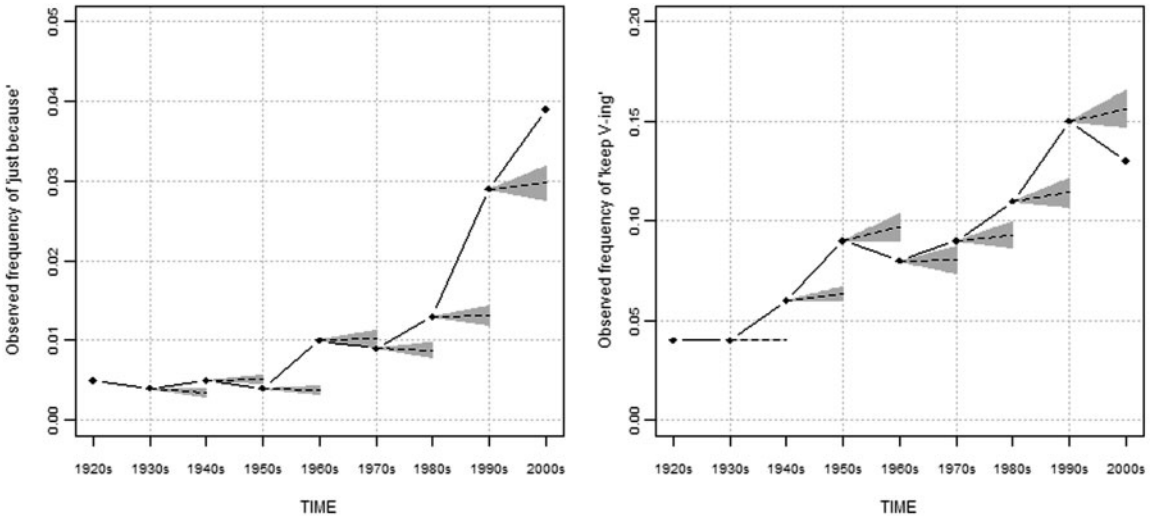

Fig. 4 Results of ISIE for *just because* and *keep* V-*ing*

In line 8, ISIE estimates an upper limit for the value of the third frequency of *just because*: it takes the value of the second frequency and adds to it the percentage sum_+ of *a*. Since this is zero that means that ISIE 'thinks' 'Since I have not observed a single increase of frequencies yet, I have no reason to expect an increase now: I therefore guess conservatively that the probable upper limit of the next frequency is the current value'.

In line 9, ISIE estimates a corresponding lower limit for the third frequency value of *just because* by taking the value of the second frequency and subtracting the percentage sum_- of *a*. Since the percentage is 1, that means that ISIE 'thinks' 'Since I have only observed a single decrease of frequencies so far, I can only assume that this development will continue: I therefore guess conservatively that the probable lower limit of the next frequency is the current value minus the last difference I witnessed'. From the upper and lower estimates, ISIE projects a range of expected values, which is shown in Fig. 4 as a gray triangle.

The interesting part comes now. As ISIE goes through additional iterations, the sum of all previous developments cumulatively shapes its expectations about future developments. Bear with us and consider the second iteration in detail: ISIE proceeds to the next frequency, 0.005 again. On this iteration, the vector of differences *d* it generates in line 2 becomes $\{-0.001, 0.001\}$, namely $0.004 - 0.005$ (from above) and $0.005 - 0.004$ (the current step). The vector *s* correspondingly becomes $\{-1, 1\}$ in line 3. The average of the absolute differences *a* remains $\{0.001\}$ in line 4. In line 5, the vector of percentages becomes $\{0.333, 0.667\}$ because the numbers from 1 to *n* (and we are looking at the second difference) get divided by the sum of the numbers from 1 to *n*. Thus, 1 and 2 get both divided by 3. In line 6, ISIE adds up all values of *p*, where *s* is positive, sum_+ becomes 0.667. In line 7, ISIE adds up all values of *p*, where *s* is negative so sum_- becomes 0.333.

On the basis of these values, ISIE estimates an expected range for the fourth corpus period. In line 8, ISIE estimates an upper limit for the fourth frequency value of *just because* as follows: it takes the value of the third frequency (0.005) and adds to it the percentage sum_+ of *a* ($0.667 \times 0.001 = 0.000667$). In line 9, ISIE estimates a corresponding lower limit by taking the value of the third frequency (0.005) and subtracting from it the percentage sum_- of *a* ($0.333 \times 0.001 = 0.000333$). Both these lines and the triangle are drawn, so ISIE guesses that the next value should be between 0.005667 and 0.004667. The actual observed

frequency lies just outside this range, as the next value is 0.004 again. In this fashion, ISIE proceeds until the triangle for the ninth value has been computed and drawn.

What is the motivation for this seemingly complex approach? First, it may be considered as an attractive complement to VNC because while the latter also provides interesting groupings of the data, these groupings are not based on all accumulated differences over time. Second, ISIE uses the mean of the absolute differences until point of time x as the overall maximal range for how the value at point of time $x+1$ may vary; thus, small ranges of variation will be predicted for curves that have not exhibited much variation in the past, which is exactly what one would want (cf. the first three prediction triangles in the left panel of Fig. 4). Third, and maybe most importantly, VNC does not take into consideration recency effects of the shape of the curve let alone weight those differentially, but ISIE does. Recall that when we looked at the third frequency, ISIE 'knew' that it was facing an uncertainty of 0.001 in the prediction of the next figure (the mean of all previous differences, a conservative estimate of the variation to expect). But then, ISIE also 'knew' that the last change it saw was an increase (from 0.004 to 0.005). Thus, while it is back at the value of the first frequency—0.005—it now recognizes an upward trend. Therefore, the uncertainty of 0.001 is not split equally between 'the next value will be larger' and 'the next value will be smaller', but since the most recent change has been an increase, that gets weighted more strongly and the triangle covers more values larger than the current one.

But what do the results actually show for *just because*? First, we obtain further support for the VNC analysis. The first four values make a very reasonable candidate for one group not only just in the VNC approach but also here, because the observed values at the beginning of the curve (where ISIE still has little information about how the data look like) are close to the predicted values (the dashed lines) and the predicted intervals (the gray triangles). The first major deviation arises after the fourth frequency, exactly where VNC also suggested a different cluster. The fifth and the sixth frequency form a group in VNC and also here since their observed frequencies are again close to the predicted values and within the gray intervals. The eighth and the ninth frequency are set apart from the rest because their observed frequencies are completely at odds with everything the algorithm could learn from the previous data and far away from the predicted values.

Second, we also get a more general feel for the predictability of the diachronic data. The data for *just because* are relatively well behaved and, within the clusters and with the exception of the last two, close to the predicted values, but the data for *keep* V-*ing* in the right panel show a rather different picture: not a single step from one frequency to another can be predicted well from what is known from the previous development. In other words, there is an overall trend but within that overall trend there are rather erratic deviations from what one would predict.

3.4 Interrupted trends

So far, we have concerned ourselves with the diachronic TIME corpus. However, as mentioned in the introduction, temporally ordered data are also available when looking at language acquisition corpora. One big difference is that because of the complex nature of language acquisition itself, the interactive spoken nature of the data, and the intricacies of collection of data in that discipline, such corpora are often not only more fine-grained but also even more variable than diachronic corpora. This is true both on the coarser level of larger trends as well as on the finer level of individual recordings. As to the former, the occurrence of utterances that are repeated several times skews frequency counts of constructions and lexical items, and the vagaries of individual recordings have a larger influence on local patterns: in a session where caretakers read to their children, the child may say less than in a session where caretakers ask a lot of questions. As to the latter, phenomena, such as U-shaped developmental patterns substantially complicate the interpretation of observed trends as do sudden turning points arising from the child having mastered a particular rule and suddenly changing the way in which a particular construction

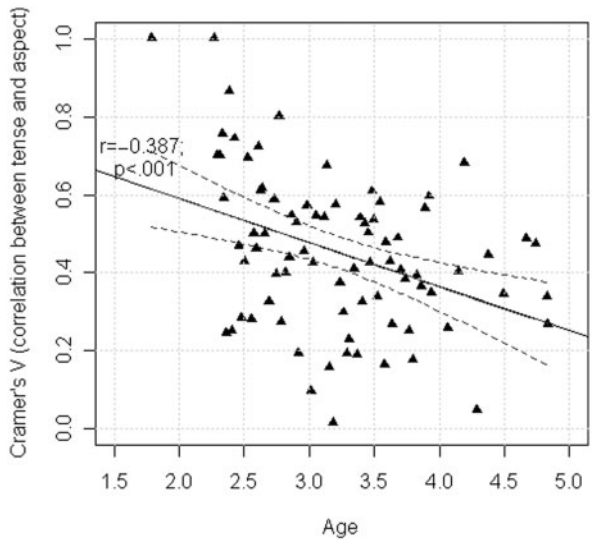

Fig. 5 Development of the coupling of tense and aspect for one Russian child

is used across the board. Some of the methods outlined above are therefore less well applicable to such data even though the general questions one wishes to investigate with the data may of course be the same.

As an example and, at the same time, a word of caution, we discuss a case study first undertaken in Stoll and Gries (manuscript under revision) concerning the acquisition of tense and aspect in Russian. Using the rather comprehensive and fine-grained Stoll corpus of Russian language acquisition, Stoll and Gries study how children gradually learn to break up the initially strong correlation between imperfective aspect and present tense on the one hand and perfective aspect and past tense on the other hand. The strength of association between the two tenses and aspects is measured using Cramer's *V*, a measure of effect size based on chi-square values. The data they obtain for one child are represented in Fig. 5, where the *x*-axis and the *y*-axis represent the temporal dimension and the strength of the tense-aspect association.

In order to determine whether there is a trend in the data, researchers could now proceed as outlined in Section 3.1 and compute a correlation coefficient. Contrary to our suggestion above, many scholars use a linear regression with a Pearson product-moment correlation. This is problematic, first, because the data violate the assumptions made by linear regressions, but even more importantly because the relation in the data need not be linear. In the present case, the data are very noisy and variable (which is why ISIE would fail here), but there is a significant correlation indicated in the graph (with a regression line and its confidence interval). However, a nonparametric smoother of the kind exemplified in Fig. 1 suggests that there are actually two trends here, not one as the simple linear regression would have us believe.

Stoll and Gries, therefore, apply an extension of regular linear regressions, regression with breakpoints. They iteratively split up the data at every individual recording into an early part and a late part and then compute linear regressions in which the dependent variable is the vector of Cramer's *V* values of the child and the independent variables are the interactions between the age and an indicator variable that marks each age as being part of the early or the late part (cf. Baayen, 2008, Section 6.4; Crawley, 2002, ch 22 for details about this approach). For each of these regressions, they stored the model deviance and then chose the

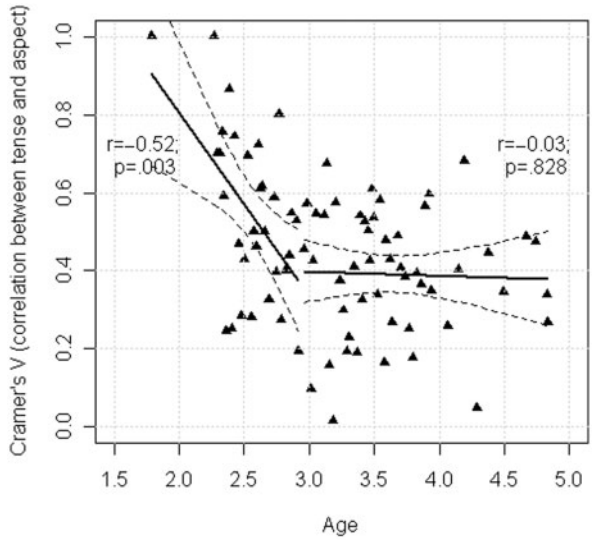

Fig. 6 Development of the coupling of tense and aspect for one Russian child 2

model whose breakpoint was smallest and after which only increasing deviances were found. The two regressions following from this—one before the breakpoint, one after it—are shown in Fig. 6.

Model comparison shows that the breakpoint at this location is highly warranted: if the amount of variance the linear model with the breakpoint explains ($R^2 = 0.266$) is compared to that of the linear model without the breakpoint from above, it emerges that the regression with a breakpoint can explain significantly more variance: $F(1, 77) = 11.991$; $P < 0.001$. In addition, the result nearly perfectly replicates the results of the smoother: from shortly before age 2 until approximately age 3, there is a strong downward trend during which the child learns to relax the coupling of tense and aspect (note the correlation coefficient, which is much smaller than the one obtained for all of the data). As of age 3, on the other hand, there is no more development and the slope of the second regression line of the child does not differ significantly from 0 anymore: no more learning with regard to the tense-aspect patterning takes place. Thus, the regression with breakpoints reveals a bifurcation of the developmental data that simple linear summary statistics and premature groupings of the data may well have missed.

4 Discussion, Concluding Remarks, and an Outlook

Frequency developments in temporally ordered corpora often present the analyst with ambiguous, unclear, or otherwise messy data. Interpreting such data on subjective grounds alone can be very problematic and may lead to incongruous conclusions. In this article, we have proposed several analytical strategies that work in a bottom-up fashion, thus allowing the data to speak for itself in a manner that may inform and guide further analysis.

In particular, we have suggested tools that address the questions whether or not the observed data reflect a genuine trend (Section 3.1) and how developments can be further partitioned to arrive at a more fine-grained understanding of the processes that have taken place. Among the analytic needs that we have addressed are the distinction of separate stages in the development of linguistic forms, even in the absence of a single linear trend (Section 3.2), the automatic detection of 'surprising' developmental breaks, given a preceding series of data points (Section 3.3), and the modeling of developments that include several linear trends that differ from one another (Section 3.4). We take it for granted

that researchers in the areas of historical linguistics and language acquisition have concerned themselves with these issues all along; our intent was to make the questions explicit and to suggest practical steps toward their solution. The tools that we have suggested in this article can not only serve to make assessments of trends more intersubjectively reliable, but we also hope to have shown that methods such as VNC and ISIE reveal aspects of linguistic changes that would not have been noticed otherwise. The potential to make previously undetected phenomena available for further analysis and, ultimately, linguistic theory building, in our view more than justifies the extra effort that comes with the application of these techniques.

We would like to conclude this article with an outlook to another type of investigation that would merit future attention. In this article, we have restricted our discussion to simple frequency counts that measure how often one or more given forms occur over a number of historical periods. Naturally, much historical research aims to go far beyond this. With regard to English, there is now a growing field of corpus-based historical sociolinguistics (Nevalainen and Raumolin-Brunberg, 1996, 2003; Reppen *et al.*, 2002) investigating the role of parameters such as gender, dialect, and genre in grammatical change. Reference to these factors can make a historical analysis not only more detailed, but can also go a long way toward explaining why a given change happened the way it did. Since these parameters are annotated in at least some of the available diachronic corpora, these can be integrated as explanatory factors in quantitative historical studies.

To take an illustrating case, Nevalainen (2000) studies the generalization of the English third-person singular present indicative suffix *-(e)s*, i.e. the gradual replacement of forms like *he hath* with *he has*. Drawing on the Corpus of Early English Correspondence, she analyzes this development not only in terms of the relative frequencies of both forms, but also in terms of regional and gendered variation, hoping to learn whether there was a particular region that initiated the change, whether one of the genders was leading the change, and how these factors interacted. Nevalainen subjects each corpus period of her data to a Varbrul analysis, determining for each one whether factors of region and gender can be used to predict usage of *-(e)s* instead of *-(e)th*.

Nevalainen's (2000) use of corpus data and her approach in general represents exactly the kind of study that could take advantage of our suggestions. Even though Nevalainen goes well beyond a simple reporting of frequencies, we see several ways in which the techniques suggested in this article could be applied to meet her analytical needs. First, there is the issue of how the corpus data are partitioned into periods. In Nevalainen's Fig. 1 and Appendix A, the data are grouped into several 20-year periods; in her Table 2 and Appendix B, there are several 40-year periods; and in Appendix C and D, there are 30-year periods representing different time slices of the corpus data with intervening intervals. A VNC approach might have been useful here to determine which groupings of the data are actually those that are supported most strongly by the data. In order to make a contrasting Varbrul analysis most revealing, we would recommend that the groupings entered as input were not arbitrarily chosen periods of time but instead periods of time that have been identified as different from each other by a data-driven, bottom-up procedure.

Second, the analysis is not as multidimensional as it could be. True, Nevalainen investigates several variables, PERIOD (not defined via a VNC), GENDER, and REGION, but as far as we can see, her analyses just involve one variable alone for each period separately (cf. her discussion on p. 48f.). A truly multifactorial approach would have used a different strategy, however. Instead of doing several different monofactorial approaches, one could use a binary logistic regression in which the dependent variable is THIRDPERSON (*-s* versus *-th*), and the independent variables are PERIOD (ordered), GENDER, and REGION, *and all the interactions involving* PERIOD:

- PERIOD and GENDER: Do the frequencies of the suffixes in question change differently over time for men and women?
- PERIOD and REGION: Do the frequencies of the suffixes in question change differently over time in the regions?

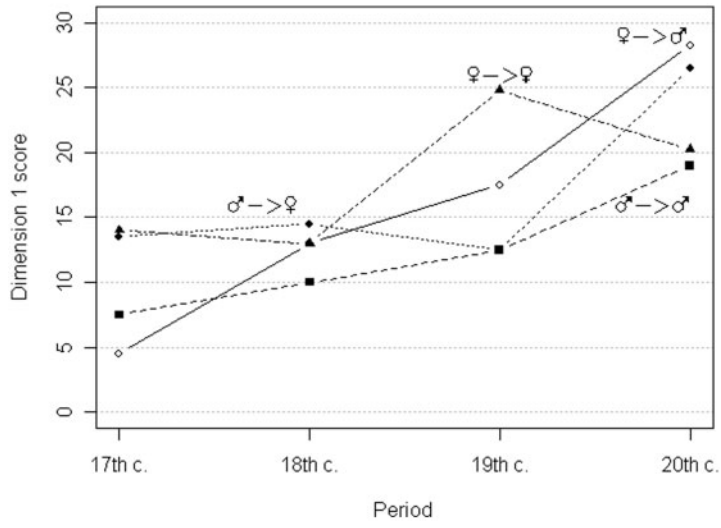

Fig. 7 Involvedness for female/male speakers/hearers (Biber and Burges, 2000, p. 31)

- PERIOD and GENDER and REGION: Do the frequencies of the suffixes in question change differently over time for the genders in the regions?

These are questions that Nevalainen's analysis does not answer, but that would be very rewarding to investigate. To illustrate this point a little further, let us consider a study that does report an interaction between the passage of time and other variables: Biber and Burges (2000) investigate the historical development of gender differences in dramatic dialogue. Their study is based on the ARCHER corpus, which they group into four periods (1650–99, 1700–99, 1800–99, and 1900–90). Each conversational turn in these texts is coded for four variables: What we can think of as the dependent variable of their study is the INVOLVEDNESS of the dramatic dialogue, which is operationalized in terms of several linguistic features that signal involved versus informational discourse (Biber, 1988). Overall, this coding produces a numerical score for each turn. The independent variables are PERIOD, SPEAKER'S GENDER, and ADDRESSEE'S GENDER. Fig. 7 summarizes the developments.

Biber and Burges observe that the sampled dialogue becomes more involved in both genders, but especially in female speakers. Interestingly, there is an interaction of PERIOD and the other variables. In the 17th century, ADDRESSEE'S GENDER shows a strong effect. Speech directed to females is more involved than speech directed to males, which tends to show characteristics of informational discourse. However, this effect wanes as time progresses. In the 20th century, the more accurate predictor of involved discourse is actually a combination of the two gender variables, namely the distinction between cross-gender and same-gender dialogues. Dialogues between the genders have become more involved, whereas this development has been comparatively weaker for dialogues for speakers of the same gender. Fig. 7 captures this interaction graphically: in the first period, the data points representing speech directed to females and the data points representing speech directed to males pattern together, respectively; in the final period, cross-gender speech and same-gender speech pattern together.

What now can our approach offer in this particular scenario? Some of the issues raised above in connection with Nevalainen's study apply here as well. VNC could be used to determine initially how the four centuries naturally partition with regard to the dependent variable of involvedness. Further, Biber and Burges do not show statistically that their conclusions from Fig. 7 can be drawn with a high degree of confidence. A multifactorial design

along the lines mentioned above could make more precise which factors are significant predictors of involvedness in a given historical period and whether that predictive power changes measurably from one period to the next. In work in progress, we are currently exploring such issues in more detail.

References

Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Beal, J., Corrigan, K., and Moisl, H. (eds) (2007). *Creating and Digitizing Language Corpora: Diachronic Databases*. Houndmills: Palgrave.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. and Burges, J. (2000). Historical change in the language use of women and men. Gender differences in dramatic dialogue. *Journal of English Linguistics*, **28**(1): 21–37.

Crawley, M. J. (2002). *Statistical Computing: An Introduction to Data Analysis using S-Plus*. Chichester: John Wiley and Sons.

Gries, St. Th. and Hilpert, M. (2008). The identification of stages in diachronic data: variability-based neighbor clustering. *Corpora*, **1**: 59–81.

Gries, St. Th. and Stoll, S. (forthcoming). Finding developmental groups in acquisition data: variability-based neighbor clustering. *Journal of Quantitative Linguistics*.

Hilpert, M. (2006). Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory*, **2**(2): 243–57.

Hinneburg, A., Mannila, H., Kaislaniemi, S., Nevalainen, T., and Raumolin-Brunberg, H. (2007). How to handle small samples: bootstrap and Bayesian methods in the analysis of linguistic change. *Literary and Linguistic Computing*, **22**: 137–50.

Kohnen, T. (2006). Historical corpus linguistics: perspectives on English diachronic corpora. *Anglistik*, **17**(2): 73–91.

Kohnen, T. (2007). Connective Profiles in the History of English Texts: Aspects of Orality and Literacy. In Lenker, U. and Meurman-Solin, A. (eds), *Clausal Connection in the History of English*. Amsterdam: Benjamins, pp. 289–308.

Leech, G. and Smith, N. (2006). Recent Grammatical Change in Written English 1961–1992: Some Preliminary Findings of a Comparison of American with British English. In Renouf, A. and Kehoe, A. (eds), *The Changing Face of Corpus Linguistics*. Amsterdam and New York: Rodopi.

Lenker, U. and Meurman-Solin, A. (eds) (2007). *Connectives in the History of English*. Amsterdam: Benjamins.

Lindquist, H. and Mair, C. (eds) (2004). *Corpus approaches to grammaticalization in English*. Amsterdam: Benjamins.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk, vol. 2: The Database*, 3rd edn. Mahwah, NJ: Lawrence Erlbaum Associates.

Mair, C. (2000). *Twentieth-Century English: History, Variation and Standardization*. Cambridge: Cambridge University Press.

Nevalainen, T. (2000). Gender differences in the evolution of standard English. Evidence from the *corpus of early English correspondence*. *Journal of English Linguistics*, **28**(1): 38–59.

Nevalainen, T. and Raumolin-Brunberg, H. (eds) (1996). Sociolinguistics and Language History. Studies based on the Corpus of Early English Correspondence. Amsterdam: Rodopi.

Nevalainen, T. and Raumolin-Brunberg, H. (2003). *Historical Socio-linguistics: Language Change in Tudor and Stuart England*. London: Pearson Education.

Reppen, R., Fitzmaurice, S., and Biber, D. (eds) (2002). *Using Corpora to Explore Linguistic Variation*. Amsterdam: Benjamins.

Stoll, S and Gries, St. Th. (in preparation). An association strength approach to characterizing development in corpora.

Notes

1 Strictly speaking, most linguistic corpora consist of texts that vary along the parameter of time. Even a corpus such as the BNC, which has been designed to represent a synchronic snapshot of British English, contains texts from different decades. Synchronic analyses abstract away from these time differences. In the end, what makes a corpus a diachronic corpus is its use for comparisons over time.

2 The *P*-values in Table 3 are exact *P*-values computed on the basis of an exhaustive permutation of all data

points; they differ slightly from what most standard statistics software would output.

3 A technical note: when VNC compares data points that arose from the merging from previous data points [as when the new data point 1965: 0.0095 is compared to the previous time period (1950) and the subsequent time period (1980), then it uses the original data points to make sure that the diversity of the data points that gave rise to the new mean is adequately reflected].