CHAPTER 10

VARIABILITY-BASED NEIGHBOR CLUSTERING

A BOTTOM-UP APPROACH TO PERIODIZATION IN HISTORICAL LINGUISTICS

STEFAN TH. GRIES AND MARTIN HILPERT

1. Introduction

Historical linguistics has been one of the linguistic subdisciplines that benefits most from corpora, and especially during the last 10 to 15 years, because many new diachronic resources have become available (Beal; Kytö and Pahta, this volume). However, the longitudinal nature and the more constrained sampling of diachronic corpora raise several problems for historical linguists. One central problem that has not been sufficiently addressed is the periodization of a linguistic phenomenon *P*, i.e. the question of how the development of *P* over time can be divided into periods or stages. This is important because neither the year-by-year development of *P* nor the predefined historical periods distinguished by corpus compilers may be particularly meaningful or revealing for *P*. It is therefore more useful for the analyst to determine the stages of *P*'s development in a bottom-up way. However, this is not easily done with existing methods. Common bottom-up methods such as cluster analysis or principal component/factor analysis cannot be readily applied to such data, as the default type of cluster analysis will produce assessments of similarity that disregard the temporal organization of the data. Hence it may group together corpus files that are in fact hundreds of years apart. This is why a different tool is needed.

In this chapter, we introduce a new clustering method to identify periods in the historical development of *P* that takes the temporal ordering of the data into consideration (Gries and Hilpert 2008; Gries and Stoll 2009). The method, VNC for "Variability-based Neighbor Clustering", involves an iterative algorithm, which in a stepwise fashion groups together those data from temporally adjacent corpus periods that are most similar to each other and, therefore, most likely to constitute a relatively homogeneous period of interest. The resulting data structure can be graphically represented in the form of a dendrogram and subjected to additional diagnostics to determine the number of periods that are most supported in the data and which data points are most likely outliers. The motivations for adopting such an approach and its application are discussed below on the basis of several case studies.

2. Motivations for VNC

It is a common practice in historical corpus linguistics to divide one's data into periods such as centuries or half-centuries. While this is rarely explicitly discussed as problematic, applying VNC instead of choosing arbitrary corpus periods has several advantages. First, a periodization into equidistant time periods may sometimes be misleading. Still, partitioning diachronic corpus data into successive stages is often necessary in order to arrive at meaningful generalizations about change. In such a scenario, it is desirable to have a method that can discern different sequential periods in a bottom-up fashion, on the basis of structures in the actual data.

The left panel of Figure 1 shows a curve representing the frequency development of the *get*-passive (Hundt 2001; Wanner 2009) in the TIME magazine corpus, which spans nine decades between the 1920s and the 2000s. The corpus contains 6,726 forms of the lemma *get* that are followed by a past participle. The graph shows a non-monotonic frequency increase: initially there is a moderate increase, followed by a plateau between 1945 and 1985, and a sharp increase after that.

A subjective description of the development might characterize the development as consisting of three phases that could be compared and analyzed further. For instance, one could ask whether the usage of the *get*-passive during its recent heyday (1990–2006) is qualitatively different from how it was used during the five previous decades of relative nondevelopment. The graph further illustrates that the

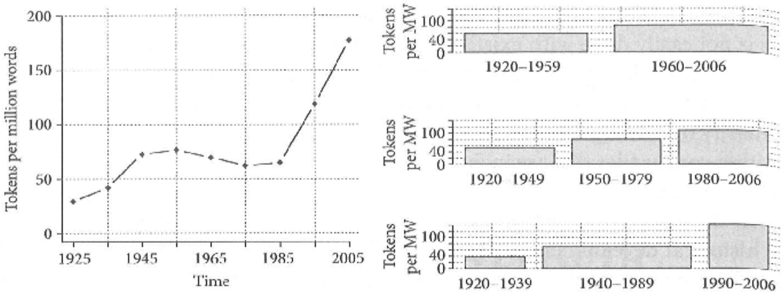

Figure 1. Frequency development of the English *get*-passive in the TIME corpus

phases into which a development is most usefully partitioned need not be equally long. Contrary to this notion, most work in diachronic corpus linguistics is carried out on the basis of equidistant periods.

The example of the *get*-passive illustrates how the choice of different periods leads to differences in the interpretation of results. The right panel of Figure 1 shows three different periodizations: if one divides the development into two half-centuries, then one characterizes the change as an increase by 78.3 percent over time (from 55.3 to 98.6 tokens pmw). However, if one divides the development into three 30-year periods, then one describes the change in terms of a 44.4 percent increase followed by a 73.1 percent increase. Finally, if one partitions based on the subjective assessment of the curve in Figure 2, one arrives at three unequally long stages with a 93.3 percent increase from stage 1 to 2 and another 113.8 percent increase from stage 2 to 3. These three perspectives are sharply different from each other and could relate very differently to explanations within an analysis.

A fair question to ask is whether the finest temporal resolution is not always the best one. Given that the TIME corpus offers the precise year in which an example was produced, why not use that information? There are different answers to this question, many of which involve a trade-off between the amount of information in the data and its interpretability. First, an analysis remaining at the level of individual data points does not leave the domain of description—generalization by means of identifying stages constitutes the first step toward explaining a phenomenon.

Second, sometimes the conflation of individual temporal data points into longer stages is necessary because the individual time points are too scarcely populated, i.e. do not contain enough data and/or enough data from different individual writers for a subsequent analysis—an appropriate conflation, however, is not marred by these problems.

Third, individual data points that are outliers (i.e. observations that behave very differently from temporally adjacent data) can make subsequent analyses very difficult because they introduce quasi-random variation into the data. An identification of stages can either be used to objectively identify such points as outliers

(and discard them) or include them in a longer stage, which reduces their potentially damaging effect on the next analytical steps. In Gries and Hilpert (2010) this has in fact made ordinal stages more predictable than interval-scaled exact years.

While these features motivate periodization, it is important to realize that what is needed is an approach that is objective (rather than based on subjective eyeballing), as well as data-driven and phenomenon-specific (rather than based on periods from theoretical accounts or different phenomena or even just convenient (half-)century splits). VNC, as a special case of hierarchical agglomerative clustering, addresses all these needs.

3. The logic of VNC

Given a range of observations (where each data point is either a single measurement or a more complex series of measurements), hierarchical clustering approaches determine which of these observations are most similar to one another and suggest groupings on the basis of mutual similarity. Since, as mentioned above, standard clustering algorithms would potentially group together temporally discontinuous observations and thus form "nonsensical" clusters, a grouping of periods is required that respects the inherent temporal sequentiality of the data. VNC is a clustering algorithm designed to operate exclusively on the variability between temporal neighbors (i.e. observations that are from successive points in time).

In this overview, we discuss two kinds of applications: first, we address cases in which a given point of time to be clustered is represented by a single numeric measurement, such as one observed frequency per point of time. (We briefly allude to how VNC can handle more complex cases.) Second, we illustrate how VNC can be used to identify outliers. Both applications have in common that the analyst has to "tell" VNC on what grounds two observations should be seen as similar. As in other clustering approaches, this is done by choosing a similarity measure and an amalgamation rule (cf. Gries 2009: section 5.5 or Hastie, Tibshirani, and Friedman 2009: section 14.3 for clarification of these terms).

As for the first choice, a similarity measure defines what counts as similarity between two observations. There are essentially two types of similarity measures: the first type, in the simplest of cases, measures differences of values, standard deviations, or (Euclidean) distances between two numerical values. The second set of measures is typically only applied to vectors of values and is based on correlations of such vectors; in the simplest of cases, one could use, say, Pearson's *r*. With such measures, two vectors of values might be seen as very similar even if the values of one are ten times as large as the other, because they share minima, maxima, etc.

As for the second choice, amalgamation rules determine how two time periods should be merged into a single one. Again, there usually is a choice between several options discussed in the above references, but we restrict ourselves here to the simplest approach of averaging values.

The VNC algorithm can be represented most generally by means of the pseudo-code in (1) below.

(1) The basic VNC algorithm in pseudo-code

```
Given a table of n temporally distinct corpus parts where each corpus part
    (i) is named by a different (average) year and
    (ii) contains numerical information about a linguistic phenomenon...
        repeat
        for any two directly adjacent corpus parts
            compute and store some measure of variability for their
            combined data
            identify the two corpus parts with the smallest measure of
            variability
            merge the data from these two corpus parts
            assign a new name to the newly merged corpus part
        until all recordings have the same name
```

VNC yields two main kinds of results, which are best represented graphically. The first is the kind of dendrogram familiar from hierarchical cluster analyses. In VNC analyses, it represents the sequence in which (neighboring) time periods were merged into clusters based on how much the individual time periods and the clusters they make up differ from each other (in units of the similarity measure chosen by the analyst). This graph can also be overlaid with the raw data to give a better impression of how these clusters map onto the observed data.

The second result is similar to a scree plot, as in a principal components analysis. It represents how much variability in the data each amalgamation step had to 'overcome' so one can identify the most useful number of clustered time periods, usually a number of periods n such that assuming $n + 1$ periods does not explain sufficiently more variability in the data.

The following two sections exemplify the logic and the results of VNC in more detail. Section 4 discusses VNC in detail as applied to points of time represented by only one value, namely the example of the *get*-passive already mentioned above; section 5 discusses VNC as a tool to detect outliers.

4. The application of VNC

To exemplify VNC with points of time characterized by single values, we return to the *get*-passives from above. Each point of time is characterized by one frequency value, which is why we use a dispersion measure (the standard deviation) as a (dis-)similarity measure: all other things being equal, the larger the standard deviation of data points, the more different they are; other dispersion measures (e.g. the variation coefficient) are also applicable. As for the amalgamation rule, we amalgamate values into clusters by averaging the individual values of time periods.

If VNC is applied to the frequency development of the *get*-passive, it proceeds as follows. It takes as input a table of corpus periods and normalized frequencies (per million words) that is shown in the top panel of Table 1.

For every sequential pair of values, the algorithm determines its standard deviation. For instance, the values of the 1920s (29.5) and 1930s (42.2) yield a standard deviation of 8.98. The two neighboring periods with the smallest standard deviation are the 1970s and 1980s (62.4 and 64.9, respectively, *sd* = 1.76). In the first iteration of the clustering process, these two are therefore merged into a single data point. This is indicated in the second panel of Table 1, which serves as the basis for the second iteration of the algorithm. This time, the 1940s and 1950s are the closest neighbors. The bottom panel of Table 1 shows this second merger. As the algorithm iterates, the time spans become larger until all nine periods are merged into a single large structure. The result of this is represented in Figure 2 as a dendrogram, overlaid with the frequencies of the *get*-passive (from Figure 1), and grey horizontal lines representing the mean frequencies of four sequential clusters.

As mentioned in the discussion of Table 1, the 1970s and 1980s are indeed merged first, and the height at which they are merged corresponds to their standard deviation (1.76). The next two periods to be merged are the 1940s and 1950s, which display the smallest standard deviation (2.75) in the second iteration of the algorithm. The first two standard deviations add up to 4.51, which is the height of the second cluster. Proceeding in this way, each cluster is merged at the height of the cumulatively summed standard deviations (cf. the left y-axis).

A given dataset can usually be grouped into different numbers of clusters. An analyst could, in principle, draw a horizontal line across Figure 2 at any

Table 1. First iterations of the VNC algorithm (based on the *get*-passive in the TIME corpus); (time periods merged in the present or a previous iteration are grey-shaded)

Decade	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s
Tokens pmw	29.5	42.2	72.8	76.7	69.6	62.4	64.9	119	177.3
Decade	1920s	1930s	1940s	1950s	1960s	1975s	1975s	1990s	2000s
Tokens pmw	29.5	42.2	72.8	76.7	69.6	62.4	64.9	119	177.3
Decade	1920s	1930s	1945s	1945s	1960s	1975s	1975s	1990s	2000s
Tokens pmw	29.5	42.2	72.8	76.7	69.6	62.4	64.9	119	177.3

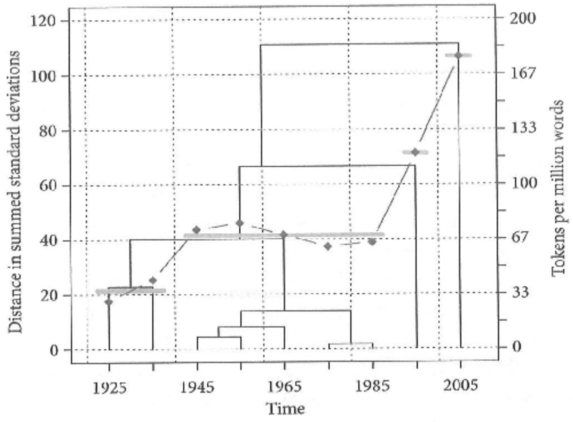

Figure 2. VNC results for the *get*-passive with overlaid frequency development

height and take the crossing vertical lines as indicating historical periods. Here, four clusters were adopted: stage 1 comprises the 1920s and 1930s; stage 2 is the frequency plateau from the 1940s to the 1980s; and stages 3 and 4 comprise the single decades of the 1990s and 2000s, respectively. Apart from the facts that this sequence agrees with an intuitive interpretation of the curve and that a sequence of four stages facilitates human interpretation, this four-cluster solution is also supported by the above-mentioned type of scree plot of the distances that the VNC algorithm measures between successively larger clusters. It was explained above how the first two periods to be merged are the 1970s and 1980s, because these are most similar. VNC then iterates and each time looks for the smallest standard deviations between data points until, in the final iteration, it merges the 2000s with the large cluster of all previous periods, which results in a large distance between the 2000s (177.3) and the previous periods is fairly large (*sd* = 44.3). The scree plot then allows one to compare the distances between successive mergers. Figure 3 plots the distances of all mergers in reverse order, starting with the last one.

Put simply, the graph shows how much dissimilarity is covered with a one-cluster solution, a two-cluster solution, a three-cluster solution, etc. The favored solution is a compromise between capturing as much dissimilarity between periods as possible and positing as few clusters as necessary. Here, the graph reveals where the first four mergers cover the lion's share of variability in the data (87.5 percent) and, thus, motivates the choice of a four-cluster solution. This can then guide the subsequent analysis of what characterizes the development of the *get*-passive across those four periods.

In the previous example, the VNC algorithm operated on the basis of one measure per individual point of time, but points of time can be characterized by more complex measures. This may be especially useful when a linguistic phenomenon does not change in frequency but in other distributional ways. While space precludes comprehensive exemplification, we want to briefly allude to how VNC handles data where each point of time is represented by a vector of values.

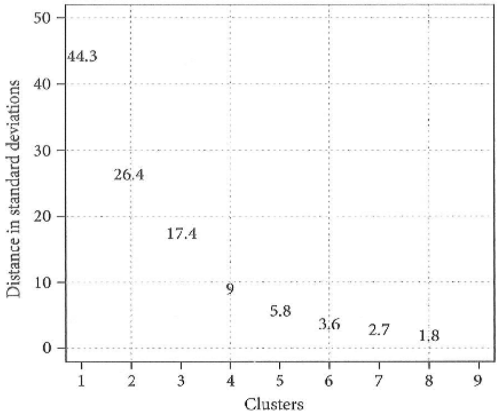

Figure 3. Scree plot of VNC results for the *get*-passive

Consider the case of the English passive (Wanner 2009). In the ARCHER 3.1 corpus, there are approximately 12,500 tokens of the construction produced by British writers between 1650 and 1989. During this time, the frequency of the construction remains fairly constant, but change in the passive can be studied, for example, by exploring how the participial collocates of the construction change over time. Consider Table 2 for an excerpt of the relevant input table, whose cells represent the frequencies of the participles (in the rows) in the time periods (in the columns).

The information in Table 2 can be used to measure similarity between adjacent periods: each period is represented by a vector of frequency values for the past participles; relative similarity between adjacent periods can be computed with a correlational statistic. The first step of VNC is to find the two periods that correlate best (i.e. whose correlation coefficient is highest). Here, periods two (1700–49) and three (1750–99) exhibit the highest correlation, so they are merged by computing the mean frequency for each past participle (1 for *abandoned*, 2 for *abated*, 1 for *abolished*, etc.). As before, this process is repeated until all time periods have been merged, and a dendrogram of the type of Figure 3 can reveal how the verb-construction associations change over time and which time periods are so different to each other that an analyst should not merge them. A more sophisticated approach would proceed not on the observed frequencies but on the verbs' collostructional preferences (cf. Gries and Stefanowitsch 2004); cf. Hilpert (2008, this volume) for discussion.

Table 2. Raw frequency data for the passive in ARCHER 3.1

	1650–99	1700–49	1750–99	1800–49	1850–99	1900–49	1950–89
abandoned	0	1	1	0	3	2	0
abased	0	0	0	1	0	0	0
abated	2	0	4	0	1	0	0
abolished	0	0	2	0	2	1	0
absorbed	0	1	0	0	0	4	4
abstracted	1	0	1	0	0	0	0
...	...	...	...	...	...	...	...

5. Detection of outliers with VNC

Another application of VNC is outlier detection, which is particularly useful for analyses involving fine-grained year-by-year data with substantial variability. For example, Gries and Hilpert (2010) study the decline of the inflectional suffix *-(e)th* in English. Between the early fifteenth century and the late seventeenth century, forms such as *giveth* were gradually replaced with *gives*. If this development is tracked in the Parsed Corpus of Early English Correspondence (PCEEC), the relative frequency of the old, interdental variant declines as shown in Figure 4.

While the overall decline is obvious, several years deviate considerably from that overall trend. If these frequencies are submitted to the VNC algorithm, the resulting dendrogram and accompanying scree plot suggest the nine-cluster solution shown in Table 3.

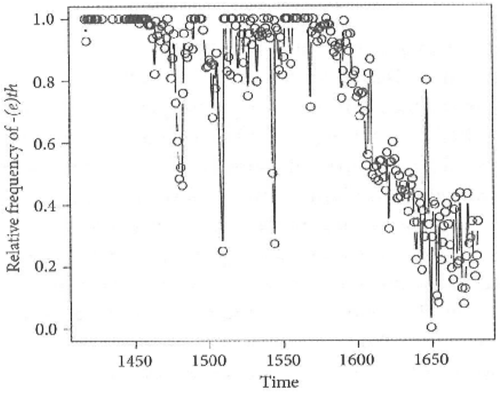

Figure 4. The relative frequency development of third person singular *-(e)th* in the PCEEC

Table 3. Initial VNC nine-cluster solution of the development from *-(e)th* to *-(e)s*

1	2	3	4	5	6	7	8	9
1417–1478	1479–1482	1483–1508	1509	1510–1543	1544	1545–1609	1610–1648	1649–1681

In Table 3, two clusters consist of only a single year, 1509 and 1544, which can also be identified easily in Figure 4. Their neighboring clusters span between 25 and 64 years, respectively. This is compelling evidence that these one-year recordings represent anomalies, and given that more than 200 data points (covering more than 250 years) enter into the analysis, the exclusion of two anomalies does not damage the substance of the database much, but excludes a huge amount of noise from the data. Gries and Hilpert (2010) thus discarded these two years from the database and re-ran the VNC algorithm, which revealed another one-year anomaly cluster (1649), which was subsequently removed from the analysis. Removal of these three data points yielded the five-cluster periodization shown in Table 4.

A remaining oddity of the data is the second period, which consists of only four years (1479–82) and shows a markedly lower frequency of the interdental variant than both of its neighbors. In Figure 4, period 2 can be identified as the short dip between 1450 and 1500 with percentages below 0.6. Period 2 is thus the only exception to an otherwise monotonic decrease of the interdental variant.

Faced with such a phenomenon, there are two options. The analyst might discard the entire cluster in order to arrive at a monotonic frequency development, or keep the period for closer scrutiny. The first option may incur the criticism of excessive data pruning; the second carries the risk that a substantial amount of noise enters the analysis. Gries and Hilpert (2010) conservatively chose to keep the second period for a subsequent analysis of the factors that bias speakers toward using either the interdental or the alveolar variant. It goes without saying that decisions of this kind must be made on a case-by-case basis, taking into account the data at hand and the planned course of analysis.

VNC thus allows the researcher to identify "bad neighbors", which can be discarded from the data intersubjectively transparently. This is an additional benefit of the methodology that makes it particularly attractive for the analysis of fine-grained diachronic data and also for data from child language corpora (cf. Gries and Stoll 2009).

Table 4. Final VNC five-cluster solution of the development from *giveth* to *gives*

1	2	3	4	5
1417–1478	1479–1482	1483–1609	1610–1647	1648–1681

6. Concluding remarks

We have described a clustering algorithm that analysts can use to partition diachronic corpus data into periods that are meaningful for the very phenomenon that they want to study. Depending on the nature of the corpus data and the phenomenon in question, different measures of similarity and different ways to amalgamate clusters can be implemented. Gries and Hilpert (2008) also discuss possibilities to extend this approach to multivariate data. The method is further useful to detect outliers. R scripts to perform any of the operations described in this chapter are available from the authors upon request, and a demo version for data of the *get*-passive type above is available on the companion website of this handbook (http://www.oup.com/us/ohhe).

References

Gries, Stefan Th. 2009. *Statistics for Linguistics with R*. Berlin: Mouton de Gruyter.

Gries, Stefan Th., and Martin Hilpert. 2008. 'The Identification of Stages in Diachronic Data: Variability-based Neighbor Clustering'. *Corpora* 3: 59–81.

———. 2010. 'Modeling Diachronic Change in the Third Person Singular: A Multifactorial, Verb- and Author-specific Exploratory Approach'. *English Language and Linguistics* 14: 293–320.

Gries, Stefan Th., and Anatol Stefanowitsch. 2004. 'Extending Collostructional Analysis: A Corpus-Based Perspective on "Alternations"'. *International Journal of Corpus Linguistics* 9: 97–129.

Gries, Stefan Th., and Sabine Stoll. 2009. 'Finding Developmental Groups in Acquisition Data: Variability-based Neighbor Clustering'. *Journal of Quantitative Linguistics* 16: 217–42.

Hastie, Trever, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. 2nd edn. New York: Springer.

Hilpert, Martin. 2008. *Germanic Future Constructions. A Usage-based Approach to Language Change*. Amsterdam: Benjamins.

Hundt, Marianne. 2001. 'What Corpora Tell Us about the Grammaticalisation of Voice in *Get*-Constructions'. *Studies in Language* 25: 49–88.

Wanner, Anja. 2009. *Deconstructing the English Passive*. Berlin: Mouton de Gruyter.

CHAPTER 11

DATA RETRIEVAL IN A DIACHRONIC CONTEXT

THE CASE OF THE HISTORICAL ENGLISH COURTROOM

DAWN ARCHER

1. Introduction

Historical courtroom data has been regularly drawn upon to investigate the diachronic development of Legal English since the 1960s (see e.g. Mellinkoff 1963) and the speech of the past more generally since the 1990s (see Doty 2010 for a more detailed discussion). This chapter argues for the sensitive use of historical courtroom data when seeking to understand the evolution of (English) legal language/the spoken language of the past. By this I mean appreciating not only the strengths and weaknesses of any primary data used but also the linguistic, social, cultural, political, and/or religious context of that data; and, hence, making extensive use of secondary sources (including the work of legal and social historians). I begin by outlining the strengths and weaknesses of some of the most popular primary sources.