CHAPTER 5

Research methods

Periodization and statistical techniques

5.1 Quantifying change

Terttu Nevalainen

5.1.1 Need for multiple methods

Past work has shown that the letter genre differs from other written genres in several respects linguistically. Analysing persistent register features in the ARCHER corpus, Biber (2001: 100) concludes that 18th-century personal letters, for example, are "nearly as involved as drama". However, in a socially stratified corpus these broad register findings do not preclude internal variation. Apart from the usual speaker variables that include age, gender and socio-economic status, issues such as the relationship between the sender and recipient of a letter have a role to play. Chapter 4 looked at the CEECE in extralinguistic terms and estimated the proportions of the various social categories sampled over time. This information provides the basis needed for interpreting in social terms the linguistic variation found in corpus data.

This chapter introduces the quantitative methods that we have adopted for analysing and comparing the processes of linguistic change investigated in this volume. Our aim is to accumulate evidence for real-time processes of change by using data sources and analytic techniques that provide a close match with past work while taking into account both group and individual patterns (cf. Tagliamonte 2012: 108–110, Brezina & Meyerhoff 2014: 23–24). However, new approaches are also called for to arrive at more adequate quantitative descriptions of real-time changes. As noted in Section 1.4, in the long term, most of the processes we analyse can be described using the S-curve model, although periods of stable variation are also found (cf. Labov 2001: 85–86). For both these alternatives, we propose a method of periodization that allows the researcher to make comparisons across the time spans of changes in the long 18th century. Section 5.1.2 discusses this approach in relation to the general issue of periodization in historical corpus studies.

The changes we study can be roughly divided into those that are based on a linguistic variable, often defined as alternative ways of "saying the same thing", and those that less easily lend themselves to a variationist approach. We adopt different

DOI 10.1075/ahs.8.c5

techniques for the quantitative analysis of these two kinds of process. Studies analysing variables are carried out using basic methods of estimating frequencies that utilize as much of the primary material as possible and allow direct comparison with past work in the field. Standard methods such as pooling frequency scores and averaging individual averages are implemented in the chapters discussing indefinite pronouns, the second-person singular *thou*, third-person neuter possessive *its*, and verbal *-s*. These methods are discussed and evaluated using some more sophisticated techniques in Section 5.2.

Quantifying processes of change that do not have a readily definable linguistic variable call for more advanced quantitative methods. We introduce and illustrate two such techniques in Section 5.3. They enable flexible sociolinguistic comparisons across different groups of people and provide an exploratory approach to the study of variation. These methods are applied in chapters that focus on periphrastic *do*, the progressive aspect and the nominal suffixes *-ness* and *-ity*. They could naturally also be applied to processes that are discussed in terms of linguistic variables and will therefore be useful in future work on such topics as well.

5.1.2 Periodizing processes of change

As shown in Chapter 4, it has been our aim as corpus compilers to facilitate the study of linguistic processes across time and the social space by presenting the material in twenty-year periods and, to enable diachronic comparisons, by using the social categories developed for earlier research on the CEEC. These sampling principles allow the researcher to compare social groups' and individual letter writers' simultaneous participation in ongoing processes of change and to identify the loci and leaders of change in social and regional terms (Nevalainen, Raumolin-Brunberg & Mannila 2011).

However, despite our efforts, the material sampled for the CEEC is unevenly distributed over time in that the corpus does not contain equal amounts of data for each social category in each 20-year period. This may become an issue with fine-grained social distinctions or shorter subperiods. Using longer periods and less fine-grained social status schemas is of course possible, depending on the research question at hand. Whatever the issue, a balance needs to be struck between data granularity and the generalizations to be made on the basis of it. Our studies therefore aggregate the data over 40-year periods as well. Similarly, a five-class model of the social order is adopted, for example, to account for the final stages of the generalization of verbal *-s* in social terms.

Some automated methods for identifying stages in the temporal dimension of linguistic changes are also available. Gries & Hilpert (2010) used variability-based neighbour clustering (VNC) to identify the temporal stages to which the diffusion

of verbal *-s* could be divided in the PCEEC data between 1417 and 1681. They arrived at five stages: 1417–1478, 1479–1482, 1483–1609 (excluding the years 1509 and 1544), 1610–1647, and 1648–1681 (excluding 1649) (Gries & Hilpert 2010: 302). This division shows that a sample bias in the corpus is replicated by a bottom-up approach of this kind. The period 1479–1482 is largely due to one family and one particular individual who contributed a large sample to the corpus but deviated from the mainstream southern English usage of the day. This bias is naturally also reflected in basic-level periodizations of the data, as in Nevalainen & Raumolin-Brunberg (2003: 68), but duly accounted for in discussions of the regional diffusion of the innovation (2003: 178). In sum, the level of abstraction often varies in periodizations. While no historical sociolinguist would set great store by a three-year period in late Middle English data as an indicator of real-time change, this brief stretch of time can nevertheless indicate significant variation in the corpus at that point.

Since linguistic changes progress at difference paces and along their unique paths over time, a common yardstick is useful in comparing processes in their own terms. In this volume we adopt a more process- and corpus-aware approach to periodization than a bottom-up analysis of the kind used by Gries & Hilpert (2010) would offer us for that purpose. To achieve that, we link the five stages of linguistic change sketched by Labov (1994: 79–83), i.e. incipient, new and vigorous, mid-range, nearing completion and completed, to the gradual diffusion of the incoming form. Labov's model reflects speakers' age levels in synchronic apparent-time analyses of phonological variables but it is also extendible to real-time analyses of processes of change. The five stages proposed in Nevalainen & Raumolin-Brunberg (2003:55) divide the slope of the real-time S-shaped curve as follows:

Incipient	below 15%
New and vigorous	between 15% and 35%
Mid-range	between 36% and 65%
Nearing completion	between 66% and 85%
Completed	over 85%

This approach allows us to compare changes and their social embedding both within and across stages. It makes it possible, for example, to compare the rate of real-time change with patterns observed in apparent time. We can ask, for example, whether it is true for real-time processes as well, as Labov (1994: 82) found to be the case for apparent time, that the rate of change is fastest at the new and vigorous stage, and slowest when the change is almost completed. Other points of comparison include the sociolinguistic patterns that characterize changes at midpoint, where we may expect contact between speakers and their individual differences to be greatest (Labov 1994: 65–66, Kurki 2005: 239–240).

Labov (1994: 82–83) further argues that in phonological changes the level of social awareness is maximal for changes nearing completion and minimal for changes at the new and vigorous stage. We can find out whether this is also the case with real-time morphological changes by comparing the stages of the changes we have analysed with comments found in normative grammars, on the one hand, and with actual usage, on the other. Another question concerning the completion of changes relates to outgoing variants and the extent to which they recede to certain limited functions or may be reappropriated in new uses.

5.2 Basic methods for estimating frequencies

Terttu Nevalainen

One of the issues with small-to-medium sized corpora is what Rissanen (1989) called "the mystery of vanishing reliability". By this telling label he referred to the problem arising with small corpora that have been divided into and annotated for various use- and user-related categories.

> The number of parameter values is, of course, inversely proportional to the amount of evidence in each information area sampled. For this reason, particularly in a corpus divided both according to chronology and text type, it may be difficult to maintain the reliability of the quantitative analysis of less frequent syntactic and lexical variants. The problem becomes even more obvious if attention is paid to sociolinguistic parameters. (Rissanen 1989: 18)

Issues also arise from the quantitative methods selected. In corpus studies, pooling subgroup data and calculating the average frequency from the total number of occurrences is commonly used in assessing both chronological developments and social variation. Subgroup data are aggregated by pooling because it is easy to carry out and provides a methodological basis for comparison with earlier studies. However, subgroups consist of individuals and, unless quota sampling is used by fixing the variable total for each individual, this method suffers from unequal contributions made by individual samples of different sizes. As Nevalainen & Raumolin-Brunberg (2003: 214–217) illustrate with CEEC data, the problem becomes acute with individuals with large samples contributing considerably more data than others and hence skewing the overall result of the analysis.

Another basic method of calculation, which avoids this problem, is computing the mean frequency of averages, that is, aggregating subgroups by calculating each individual's average frequency separately and then calculating the average of these

averages. This way each informant is given equal weight in the outcome. But sample size and the number of attestations may again affect the outcome: if one writer has one instance of the incoming form and none of the outgoing one, the contribution of this individual would be considered equally important as that of another writer who produced no instances of the new form but as many as 40 of the old one. Using this method, we will need to set minimum frequencies for the number of observations for each informant to prevent such extreme disparities.

An issue related to both these methods, pooling and averaging of averages, is poor *dispersion*, a situation in which there is only a small number of informants, who show large differences in the amounts data they have produced. Hence some minimum frequencies for the number of contributing individuals will also need to be set. Ideally, adequate frequency levels can be set by testing how well subgroups match the larger datasets that they form part of.

Pooling and averaging averages are both maximum likelihood methods in that they try to obtain one estimate for the value of the unknown frequency of the incoming form. There are other, more sophisticated methods that can be used for estimating frequencies. In earlier work we compared pooling and averaging of averages with both *bootstrapping* and the *Bayesian approach* (Hinneburg et al. 2007, Mannila, Nevalainen & Raumolin-Brunberg 2013). In bootstrapping, the uncertainty in the frequency estimate is examined by resampling the original data a large number of times (cf. also permutation testing discussed in Section 5.3.2). For each bootstrap sample a method such as pooling can be used for estimating the frequency of the incoming form, and the median of these estimates is then taken to represent the bootstrap estimate for the frequency of the incoming form. The Bayesian method is rather more complicated taking into account both the population frequency and each individual's frequency in order to produce probable values for the parameter studied.

One way to assess the degree to which the simple and more sophisticated methods converge is to apply them to the same data set. Figure 5.1 compares the frequency estimates obtained using all four methods to study the gradual replacement of the third-person singular suffix *-th* by *-s* in *have* (*hath* v. *has*) between 1580 and 1660. The calculations are based on a dataset consisting of over 400 people and the total of 2,464 instances of *hath* and 472 of *has*. The four curves all indicate the relatively slow process made by *has* in the 80-year period, basically suggesting that it progressed from the incipient to the new and vigorous stage.

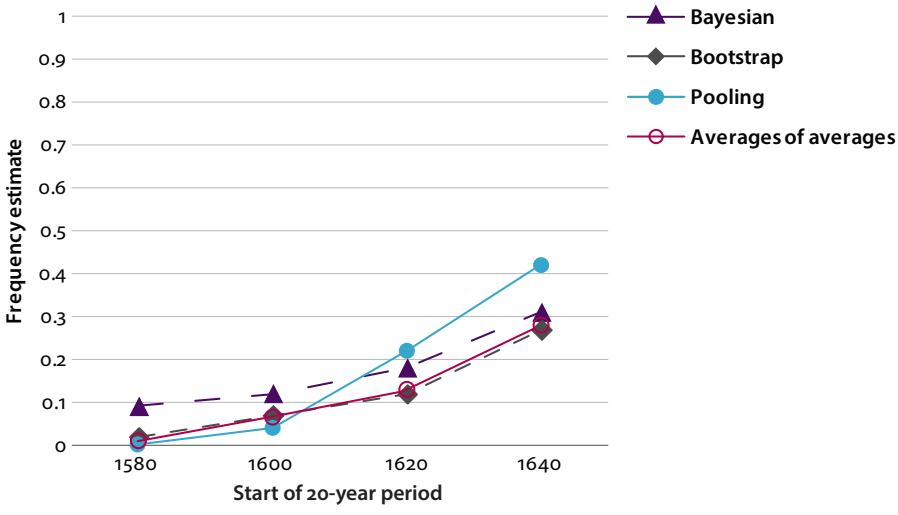

Figure 5.1 Results given by four estimation methods for the frequency of *has* in four 20-year periods from 1580 to 1660 in the CEEC

All four methods produce quite similar results with the exception, from the 1620s onwards, of pooling, which indicates higher frequencies than the other three. Because of the prior information required by the Bayesian method, the first period starts on a higher level with that technique than with the others. Averaging of averages and bootstrapping produce closely matching results. This was the general conclusion that we came to applying these four methods to different datasets in Mannila, Nevalainen & Raumolin-Brunberg (2013).

Unlike pooling and averaging of averages, the bootstrap and Bayesian methods produce intervals representing the degree of uncertainty in the estimates. With bootstrap methods, the standard deviation of sample frequencies can be used to yield a confidence interval for the frequency in the original data. These confidence intervals directly reflect the amount of data analysed. For the data in Figure 5.1 the confidence intervals were very narrow and for the bootstrap method, for example, overlapped only partly in the first two periods and not at all in the last two. In Mannila, Nevalainen & Raumolin-Brunberg (2013) we found that having at least 15 persons with at least 10 instances per variable in a bootstrap estimate gave a good fit between a subgroup and the full dataset from which it was drawn. Even fewer instances, such as five per person, could give a reasonable fit, provided that the number of informants was large enough.

However, it is a common occurrence that the researcher is left with smaller subgroups especially with low-frequency linguistic features or when studying short time periods or a range of socio-economic categories. This was the case with the

analysis we carried out to determine whether individuals were progressive or conservative with respect to an ongoing change in Nevalainen, Raumolin-Brunberg & Mannila (2011), where we applied the bootstrap method to the CEEC data in a number of linguistic changes. To minimize skewing, we only examined periods for which there was data on the use of the study variable for at least five individuals. Moreover, the procedure for determining whether an individual was progressive or conservative was applied only if the person had at least six occurrences of the variable, that is, the sum of the occurrences of the outgoing and the incoming variant for the person was at least six.[1]

In Part II, the variation studies of linguistic changes make use of the basic methods discussed above, bearing in mind their limitations especially with low-frequency variables. Some chapters also consider the differences produced by pooling and averaging of averages, and those that analyse linguistic variables use bootstrapping to assess the degree of uncertainty in their estimates. Most of the studies also raise the level of abstraction by aggregating data over longer periods and over broader social categories than those specified in our sampling frame. These measures help diminish the degree of "vanishing reliability" and make for more reliable results. The measures used naturally impact on the generalizations that can be made on the basis of the quantitative findings. Using multiple methods, we hope to offer empirical baseline information of different kinds on the sociolinguistic contexts of language change in 18th-century English.[2]

Finally, recognizing the internal heterogeneity of subgroups, we also discuss outliers, individuals who deviate from others, either leading the process of change or lagging considerably behind their contemporaries. They are of particular interest both as unique individuals and as representatives of intersecting sociolinguistic categories and communities. Since we are always analysing the same set of people, the ways in which these individuals pattern with respect to linguistic changes in their different stages make interesting comparisons. This issue is addressed with all the processes studied, regardless of whether they can be construed in basic variationist terms or not.

---

1. For further discussion of these and other quantitative methods applied to the CEEC, see Hinneburg et al. (2007), Nevalainen, Raumolin-Brunberg & Mannila (2011), Mannila, Nevalainen & Raumolin-Brunberg (2013) and Nevalainen (2014a).

2. Nevalainen & Raumolin-Brunberg (2003: 193–200) used a Variable Rule Analysis application (GoldVarb) to study the relative impact of the various social variables on processes of change. In this volume it would have been possible to update the traditional Varbrul toolkit and resort to more recent techniques such as Rbrul (e.g. Tagliamonte 2012: 138–157). We decided not to pursue that line of inquiry but to explore a variety of measures and make more transparent the basic principles of counting frequencies instead.