Trim: 247mm × 174mm Top: 10.5mm Gutter: 22.001mm
CUUK2979-02 CUUK2979/Kytö & Pahta ISBN: 978 1 107 03935 3 December 8, 2015 16:16

# 2

# Quantitative approaches to diachronic corpus linguistics

Martin Hilpert and Stefan Th. Gries

## 2.1 Introduction

English historical linguistics has a rich and long-standing tradition of corpus-based work (see the surveys in Rissanen 2008, Kytö 2012). Resources such as the Helsinki Corpus, the Brown family of corpora and ARCHER have spawned active research programmes for the study of lexical and grammatical change, both long term (Curzan 2008) and short term (Mair 2008). In addition, corpus resources inform the analysis of diachronic variation in genres (Hundt and Mair 1999), registers (Biber and Gray 2011b) and varieties (Tagliamonte 2006b). The present chapter will discuss a currently developing line of research which uses the methods of *quantitative* corpus linguistics for the analysis of *diachronic* corpora. This research program draws on, and is informed by, the aforementioned areas, but at the same time, it uses particular kinds of data and handles that data in specific ways that merit discussion. Diachronic corpora are understood here as textual resources that represent comparable types of language use over sequential periods of time, thus comprising at least two periods, as in the Diachronic Corpus of Present-Day Spoken English (DCPSE, Wallis et al. 2006), but typically many more, as in the Corpus of Historical American English (COHA, Davies 2010), a monitor corpus which at the time of writing samples twenty-one sequential decades of language use (see Chapter 8 by López-Couso in this volume). The English diachronic corpora that are currently available represent different varieties and text types and vary in their respective time depths, but it is a design feature of most diachronic corpora to hold the type of text constant, so that diachronic language change within a given text type may be studied with as few confounding factors as possible. Quantitative corpus linguistics (Biber and Jones 2008) is a research tradition in which research questions are formulated in such a way that frequency counts from corpora may provide answers. Quantitative corpus work thus often engages in hypothesis testing, so that a testable empirical question (e.g. 'Have adolescent women been leading the development of the quotative *be like* in Tyneside English?')

Trim: 247mm × 174mm Top: 10.5mm Gutter: 22.001mm
CUUK2979-02 CUUK2979/Kytö & Pahta ISBN: 978 1 107 03935 3 December 8, 2015 16:16



may receive an answer in terms of either 'yes' or 'no'. Of at least equal importance are so-called exploratory techniques, which are designed to transform a complex dataset into a summary (and often visual) representation (which may then be interpreted by the analyst and that may in turn lead to the formulation of hypotheses). To give an example, Szmrecsanyi (2010) studies the use of genitive constructions in different text types of British and American English in the 1960s and the 1990s, exploring whether there are changes that could be seen as Americanization or colloquialization (see Mair 2006). The frequency counts that enter quantitative corpus studies often represent token frequencies, but a much wider variety of measures is routinely used, including measures of type frequency, dispersion, and collocation.

The main point of this chapter will be an overview of how the two, diachronic corpora and quantitative corpus linguistics, are put together in fruitful ways. Quantitative studies of how units of linguistic structure change across corpus periods can address questions of more general linguistic interest, including the following:

- When and how does a given change happen?
- Can a process of change be broken down into separate phases?
- Do formal and functional characteristics of a linguistic form change in lock-step or independently from one another?
- What are the factors that drive a change, what is their relative importance, and how do they change over time?
- How do cases of language variation in the past compare to variation in the present?

It is already apparent from these questions that quantitative studies of historical change have a great deal in common with quantitative studies of synchronic variation (Tagliamonte 2006a), both on the theoretical and the methodological level. This commonality is of course no coincidence, as language variation is one key factor for explaining why languages change over time. The remainder of this chapter is organized in the following way. Section 2.2 motivates the approach that is taken here and explains how quantitative methods usefully complement qualitative approaches in the analysis of diachronic corpus data. Section 2.3 is concerned with approaches to the diachrony of variation in language, and it discusses desiderata of such approaches. Section 2.4 turns to exploratory techniques, which can guide the researcher towards discovering new, unanticipated aspects of language change or assist in the formulation of hypotheses. Section 2.5 offers a few pointers for future research and section 2.6 concludes.

## 2.2 Language change by the numbers

Historical linguistics, by its very nature, depends on the observation of authentic data. However, not all research questions in historical linguistics

Trim: 247mm × 174mm Top: 10.5mm Gutter: 22.001mm
CUUK2979-02 CUUK2979/Kytö & Pahta ISBN: 978 1 107 03935 3 December 8, 2015 16:16



oblige the analyst to quantify that data. Many processes of linguistic change manifest themselves in qualitative differences, so that for instance lexical items disappear from usage, or word order patterns that once were common are no longer used. (Of course, such differences can be quantified as observed frequencies becoming zero.) For instance, the Old English (OE) word order shown in (1), an example from Ælfric's *Homilies of the Anglo-Saxon Church* (ÆC Hom I, 1.20.1), is no longer used in Present-day English (PDE).

(1) on twam þingum hæfde God þæs mannes sawle gegodod
in two things had God the man's soul endowed
'God had endowed man's soul with two things'

The crucial characteristic of the example is the fact that the finite verb *hæfde* 'had' appears after an initial constituent in what is called the 'verb-second' position (Fischer et al. 2000). A gloss such as *With two things had God man's soul endowed*, which retains this particular word order, might be acceptable as a deliberate anachronism, but it will not pass as an everyday PDE sentence. Hence in this case, a single historical example, in connection with the intuitions of a present-day speaker, is enough to establish that a change has taken place.

However, more rigorous quantification of diachronic data becomes necessary when research questions go beyond the mere detection of a change and into the internal dynamics of that change. This means that, often, approaches are required that meet the following criteria:

- They are *multifactorial* in that they take multiple formal, functional and language-external/social features, or predictors, causes for linguistic choices into consideration rather than just comparing counts of the results of linguistic choices.
- They involve *interactions* between the formal, functional, and language-external/social predictors so that one can determine whether a particular predictor has the same effect regardless of other predictors' values. While most studies simply adopt the assumption that the effects of different predictors hold independently from one another, this need not be the case, and one can only identify such cases when tests for interactions are included.
- They involve *interactions of*, say, *Time (or Corpus)* on the one hand and formal, functional and language-external/social predictors on the other hand so that one can determine whether a predictor has the same effect in each time period or whether the role a particular feature plays for speakers' choices changes over time. Without such interactions, it is nearly impossible to make principled comparisons between different time periods.

Some studies already involve these more sophisticated approaches, usually in the form of multifactorial regression analyses. Such regression analyses

Trim: 247mm × 174mm Top: 10.5mm Gutter: 22.001mm
CUUK2979-02 CUUK2979/Kytö & Pahta ISBN: 978 1 107 03935 3 December 8, 2015 16:16



try to predict the outcome of a dependent variable (or response) on the basis of one or more independent variable(s) (or predictors). Crucially, both the response and the predictors can be of different kinds, i.e., they can be binary (ditransitive vs. prepositional dative), categorical and/or ordinal (human vs. animate vs. inanimate vs. abstract), or numeric (time or length of a word in phonemes); depending on the nature of the dependent variable, one would use binary logistic regression, multinomial or ordinal logistic regression, or linear regression. Also, a central advantage of these regression models is that they allow the researcher to study the effects of several predictors (and their interactions) at the same time (Baayen 2008: chs. 6–7; Gries 2013: ch. 5) so that researchers can determine which predictors affect linguistic choices significantly, in which direction (does a particular predictor make a choice more or less likely), and how strongly. In spite of these many advantages of regression modelling, there are still many studies that do not involve the proper comparisons of observed frequencies of phenomenon P in different time periods (see Gries 2011 for discussion of an example).

Returning to our verb-second example from above, this means that if we want to find out how verb-second word order gave way to the patterns that are in use today, neither looking at individual examples nor mere tabulated frequencies of verb-second and other orders are sufficient. Rather, we need to identify the contexts in which verb-second disappeared first, and we would need to identify the formal, functional and language-external/social features that characterize these contexts. On an abstract level, the answers that one is usually looking for derive from all three above criteria: one wants to be able to indicate that 'during time period X, context feature Y biased speakers towards the new, incoming word order pattern with a relative strength of Z' (Hilpert 2013: 50). By analysing the impact of a range of context features over a range of time periods, we thus arrive at a differentiated picture of how the change in question proceeded. Most importantly, we learn which contextual features play an important role and which ones do not, and we can find out whether the effects of these features change in strength over time. We might also find that two contextual features interact in such a way that, for instance, they only have an effect if they co-occur, but not if they occur in isolation. Observations of these kinds are difficult, if not impossible, to make on the basis of individual examples; quantitative corpus analysis thus works like a magnifying glass, allowing the researcher to detect phenomena that would not otherwise be open to inspection.

It is important to realize that this higher level of observational detail is no end in itself: having precise information about how a given change happened is a necessary prerequisite for discussions of why the change happened in the way it did. Are we looking at a change that can be connected to social developments (Americanization, colloquialization), do the data support the idea of culture-bound, genre-specific developments (complexification, simplification), or can the change receive a structurally motivated explanation (generalization, analogical levelling)? Claims that link observations of change to

Trim: 247mm × 174mm Top: 10.5mm Gutter: 22.001mm
CUUK2979-02 CUUK2979/Kytö & Pahta ISBN: 978 1 107 03935 3 December 8, 2015 16:16



these potentially competing motivations of change must be based on analyses in which alternative explanations are considered with due diligence. It is here that quantitative techniques have a decisive advantage over qualitative assessments of change: a quantitative analysis can simultaneously weigh the relative impacts of several factors, thus separating the wheat from the chaff. The analysis may for instance demonstrate that a given factor only has a very small effect, or even no effect at all, so that explanations related to that factor can be ruled out – at least for the sample that is being analysed and the population for which it is representative. Demonstrating this on the basis of qualitative data, in a way that will convince a sceptical reader, is a very difficult task. While it goes without saying that any quantitative study is of course grounded in a fundament of qualitative insights, it should equally go without saying that the analysis of language change by the numbers is an indispensable tool for extending those insights. The next two sections flesh out this statement with a number of concrete examples.

## 2.3 Quantitative analyses of diachronic variation in language

How does language variation in PDE compare to variation in earlier periods of English? In a study that addresses this question, Wolk et al. (2013) use the ARCHER corpus to investigate how variation in genitive and dative constructions has changed during Late Modern English. What these constructions have in common is that they are organized in paradigmatically related pairs, so-called alternations. The member constructions of an alternation are available as alternative ways of verbalizing the same, or at least fairly similar, conceptual content. Examples of the genitive alternation and the dative alternation are shown in (2).

(2a) the prince's horse the horse of the prince
(2b) I wrote him an email I wrote an email to him

Synchronic analyses of both the genitive alternation (Hinrichs and Szmrecsanyi 2007) and the dative alternation (Bresnan et al. 2007) have identified several factors that probabilistically affect speakers' choices between the respective alternative constructions. Those factors include semantic characteristics such as animacy, pragmatic characteristics such as the topicality/givenness of referents, and formal characteristics such as definiteness, pronominality, or length of (the referents of) possessors/possessees and recipients/patients. To illustrate the workings of just one factor with regard to the genitive alternation, the *s*-genitive construction is relatively less tolerant towards inanimate possessors than the *of*-genitive construction (?*the water's temperature* vs. *the temperature of the water*). In the dative alternation, the prepositional dative construction is relatively more tolerant towards syntactically heavy constituents in the recipient slot (?*I wrote my sister, who lives in Spain,*

Trim: 247mm × 174mm Top: 10.5mm Gutter: 22.001mm
CUUK2979-02 CUUK2979/Kytö & Pahta ISBN: 978 1 107 03935 3 December 8, 2015 16:16



*an email* vs. *I wrote an email to my sister, who lives in Spain*). Experimental studies show that PDE speakers have internalized the complex ecologies of the determining factors in these alternations (Bresnan 2007), but it stands to reason that, historically, there must have been developments leading up to the status quo. The exemplary study of Wolk et al. (2013) is a case where researchers aim to determine how these developments unfolded.

For each of the two alternations, Wolk et al. (2013) determine a variable context and retrieve all relevant examples from the ARCHER corpus. Each example is annotated in terms of a dependent variable, which marks the respective constructional choice, and in terms of several independent/explanatory variables, or predictors, such as animacy, topicality/givenness, definiteness, and crucially also the historical time period during which the example was produced. Wolk et al. (2013) then perform binary logistic regression analyses in order to obtain results that can be compared against earlier studies that analysed synchronic data, and that also indicate whether the impact of those factors has become weaker or stronger over time. Overall, the results that Wolk et al. (2013) obtain reaffirm findings based on synchronic data. The factors that are analysed show effects in the expected directions, which allows the conclusion that there has been substantial diachronic stability in the use of both genitive and dative constructions.

However, there have also been changes. For the genitive alternation, Wolk et al. (2013) find a diachronic change in the effect of the length of the possessed entity. Generally and following from a general short-before-long tendency in English, a longer possessed entity favours the *s*-genitive (*John's sixteen-year-old stationwagon*), but Wolk et al. (2013) note that the relation is non-linear when it comes to short possessed entities, especially in their early corpus data. Very short possessed entities are thus not necessarily strongly drawn towards the *of*-genitive. Over time, that non-linear relation becomes more linear: in the words of Wolk et al. (2013), length is '*more well-behaved*' in later corpus periods. A second change involves the semantic factor of animacy. Whereas *s*-genitives in eighteenth- and nineteenth-century English rarely occurred with collective, locative, or temporal possessors (e.g., *the Academy's decision*, *the island's inhabitants*, *today's technology*), the frequencies of these options sharply increase in the twentieth century, pointing to a process of semantic generalization. As for the diachronic development of the dative alternation, animacy is also shown to play a role. Inanimate recipients, as in *The herbs gave the soup a nice flavour*, have become more frequent in the ditransitive construction in the twentieth century.

These findings demonstrate that the probabilistic usage patterns of constructions undergo fine-grained changes that could not be detected through the comparison of individual examples, but that do lend themselves to meaningful interpretations in terms of general processes of language change.

In another study that targets change in variation, Buchstaller (2011) investigates quotation markers in Tyneside speech on the basis of a tripartite corpus that consists of sociolinguistic interviews collected in the 1960s, the

Trim: 247mm × 174mm Top: 10.5mm Gutter: 22.001mm
CUUK2979-02 CUUK2979/Kytö & Pahta ISBN: 978 1 107 03935 3 December 8, 2015 16:16



1990s, and the late 2000s. Whereas the variation of genitive or dative constructions involves only two alternative expressions, so that the dependent variable has only two levels, matters are a little more complex in the case of quotation markers. Here, speakers can draw on a set of several forms, and the recent addition of innovative variants such as *go* or *be like* suggests that the system of quotative markers is currently undergoing a substantial reorganization. Examples of some of the variants in Buchstaller's data (2011: 59) are given in (3).

(3a) I never say 'howay man'
(3b) I shouted back 'well if you stop kicking the door . . . '
(3c) I just went up to him and Ø 'excuse me mister . . . '
(3d) She was like 'eeh! It's a rodent!'
(3e) She goes 'I might not wear them'
(3f) I'm all, 'Dude, you're not helping your cause!'

Buchstaller (2011) sets out to investigate diachronic changes in the extralinguistic and intralinguistic factors that influence speakers' choices in that system of variants. Again, research on synchronic variation (Buchstaller and D'Arcy 2009) has identified several determining factors, such as the content of the quote, i.e. whether a thought, utterance, or noise is quoted, the grammatical tense that frames the quotation marker, the grammatical person of the quoted speaker, the distinction between narrative and other texts, and social variables such as age, social class and gender. Buchstaller (2011) exhaustively retrieves examples of quotation from her corpus. The results identify *say* as the most frequent variant throughout, which however decreases over time in relative frequency with the emergence of *go* and *be like*. But how are these frequency developments reflected in a changing ecology of determining factors? In order to approach this issue, Buchstaller (2011) first examines each factor on its own.

As for the extralinguistic factors, the quotative system in the 1960s is differentiated by gender, but not by age or social class. This subsequently changes: with *go* becoming more frequent in the 1990s, and *be like* even surpassing it in frequency in the 2000s, age and class, in addition to gender, become relevant determinants. Young women are the speakers that adopt *be like* to the greatest extent. As for the intralinguistic factors of quotation content, grammatical person, and grammatical tense, these exert an influence throughout the three corpus periods, but patterns of change emerge here, too. For instance, whereas *say* is the preferred marker of first-person quotations in the 1960s, it has ceded that role to *be like* in the 2000s, during which *say* and *go* show an inclination towards third-person quotations.

With a complex dataset that reflects several factors influencing speakers' choices between several forms, the analyst has to rely on multifactorial/-variate statistics to arrive at reliable generalizations. Buchstaller (2011) performs a multinomial regression analysis (Gries 2013: ch. 5) of the complete

Trim: 247mm × 174mm Top: 10.5mm Gutter: 22.001mm
CUUK2979-02 CUUK2979/Kytö & Pahta ISBN: 978 1 107 03935 3 December 8, 2015 16:16



dataset, which yields that the effects of age, social class, grammatical tense, and narrative are measurably different across the three subcorpora. In other words, the emergence of new variants in the quotative system of Tyneside English goes along with a reorganization of the selection processes that speakers of different age groups and different social classes make. Unlike the system of genitive and dative constructions, which undergoes just minor rearrangements, the results show that the system of English quotation markers is currently in a state of upheaval that might either stabilize or see further change through the repeated intrusion of new variants. Buchstaller's quantitative analysis (2011) pinpoints the exact loci of change and indicates what factors change at what time. It thus gives an affirming answer to the question whether young women have been spearheading the emergence of quotative *be like*, but at the same time, the results offer a picture that is much more differentiated than that.

It was mentioned in the introduction of this chapter that quantitative corpus-based methods are commonly applied in order to test hypotheses. Whereas the two case studies that were described above address fairly specific research question, their primary aim was not to decide between two rivalling hypotheses. A study that checks the validity of a pre-existing hypothesis is presented by Geeraerts et al. (2011), who investigate the emergence of *anger* as a term that ousted its near-synonyms *ire* and *wrath* during Middle English. Diller (1994a) suggests a socially motivated explanation for this development, hypothesizing that *anger* emerged as an expression for annoyance in lower-ranked persons, as opposed to the *ire* and *wrath* of socially powerful beings such as kings or deities. From this hypothesis, Geeraerts et al. (2011) derive the predictions that *anger* should be used to describe situations in which the social status of the experiencer is low, the offense affects only the experiencer, rather than having more profound consequences, and the experiencer's reaction to the offense is non-violent. Geeraerts et al. (2011) retrieve all tokens of *ire*, *wrath*, and *anger* from a collection of Middle English text and annotate those in terms of the semantic factors outlined above, as well as distinguishing between tokens from religious and non-religious text and between translated and natively produced texts. The analysis further includes historical time as a variable, distinguishing examples from approximately 1300, 1400, and 1500. Analyses of each individual semantic variable across those three time periods reveal processes of change for the social status of the experiencer and the affectedness of the experiencer, but not for the violence or non-violence of the reaction to the offense. Geeraerts et al. (2011) then use a binary logistic regression (Gries 2013: ch. 5) to assess the combined effects of the described factors, the dependent variable is modelled as a contrast between *anger* and the combined tokens of *ire* and *wrath*. The results are largely in line with Diller's hypothesis (1994a). The use of *anger* at 1400 is favoured by contexts of personal offences with non-violent reactions, a marginally significant effect is observed for low social ranks

Trim: 247mm × 174mm Top: 10.5mm Gutter: 22.001mm
CUUK2979-02 CUUK2979/Kytö & Pahta ISBN: 978 1 107 03935 3 December 8, 2015 16:16



of the experiencer. The effect of non-violent reactions is stronger in non-religious texts than in religious texts. The data further show that over time, as *anger* becomes the default term for the emotion it denotes, these effects weaken. Examples from around 1500 thus have a relatively higher likelihood than earlier ones to denote public offences of high-ranking experiencers that react violently to those offenses.

In summary, the case studies presented in this section illustrate three issues. First, the variationist approach to analysing the use of alternative expressions with similar functions is fruitfully transferred to the usually regression-based analysis of variation over historical time. With a diachronic corpus that represents sequential periods of English, time can be included into the analysis as one (interacting) predictor among others, and it can be determined how variation in the present compares to variation in the past. Second, this type of analysis offers nuanced accounts of what has happened, so that it can be specified what factors had an effect at what time. The contrast between the studies by Wolk et al. (2013) and Buchstaller (2011) shows that the dynamics of diachronic variation may range from relative stability to substantial reorganization, which requires a fine level of observational granularity: an analysis has to do more than just ask whether or not a particular factor has an effect – it has to ask when this effect obtained and how it varied in strength over time. Third, the observations that these studies offer importantly include the absence of effects, which is evidence that can in principle serve to rule out hypotheses that predict those effects. An aspect that has not received much attention in the discussion above is that the findings from quantitative studies usefully feed back into the development of linguistic theories, either enriching already existing theoretical claims or generating altogether new hypotheses. The idea of using quantitative corpus-based methods to generate new ideas is taken up more extensively in the following section.

## 2.4 Quantitative analyses of diachronic change: exploratory approaches

An attractive potential of quantitative corpus-based methods that has yet to be fully realized in diachronic studies lies in exploratory, bottom-up approaches (Gries 2011). The label 'bottom-up' stands for a set of techniques in which the data are processed statistically in order to discover structures that had not necessarily been anticipated by the analyst. Compared to the approaches that were presented in the previous section, these methods often reverse the order of qualitative and quantitative analysis. Whereas for instance a logistic regression analysis requires a fundament of qualitative analysis which is subsequently scrutinized statistically, bottom-up approaches may start with the statistical processing of raw data, which then yields results that function as a stepping stone for a qualitative analysis.

Trim: 247mm × 174mm Top: 10.5mm Gutter: 22.001mm
CUUK2979-02 CUUK2979/Kytö & Pahta ISBN: 978 1 107 03935 3 December 8, 2015 16:16



Starting with automated computational procedures has the benefit of a 'fresh start' that may serve to eliminate preconceptions and to reveal previously overlooked aspects of a given phenomenon.

One example for such an approach is Sagi et al. (2011), who apply a bottom-up computational approach to the study of lexical semantic change. Whereas word meaning is usually thought of as an area of study in which the intuitions of a human analyst are completely indispensable, research in natural language processing has developed a range of methods that operationalize the meaning of a given word in terms of the elements and structures that occur in the linguistic context of that word. J. R. Firth's dictum that '*You shall know a word by the company it keeps*' (1957: 11) has thus found its way into methods such as latent semantic analysis (Landauer et al. 1998), which produce results that stand up to comparisons with human processing of word meaning. Latent semantic analysis uses corpus data to characterize word types in terms of frequency lists of their collocates. For instance, the noun *toast* frequently occurs close to nouns such as *tea*, *cheese*, *slice*, and *coffee*. A statistically processed frequency list of all collocates of *toast* is called its semantic vector. Semantic analysis enters the picture when semantic vectors of several words are compared. Two words are in a semantic relation if their semantic vectors are highly similar. For instance, near-synonyms such as *cup* and *mug* will have similar semantic vectors, but also converses such as *doctor* and *patient* and even antonyms such as *hot* and *cold*. If a large group of semantic vectors is analysed with a dimension-reducing technique such as multidimensional scaling (Wheeler 2005) or correspondence analysis (Greenacre 2007), semantic relations between those words can be visualized in two-/three-dimensional graphs in which words with close semantic ties are positioned in close proximity whereas semantically unrelated words are placed further apart.

Whereas most applications of latent semantic analysis analyse word types, thus averaging collocate frequencies over many occurrences of the same word, Sagi et al. (2011) use an approach that operates at the level of word tokens, thus capturing meaning differences between individual occurrences of the same word. In order to overcome data sparsity, that method uses not only the direct collocates of the target word, but also second-order collocates, that is, the collocates of collocates. Given a concordance line such as *he prescribed tea and toast and a small bit of steak*, the second-order collocates would include the word *doctor* (a collocate of *prescribed*) and *coffee* (a collocate of *tea*). The latter will be relatively more important, since it is also a collocate of *toast* itself. Applied in this way, latent semantic analysis can transform a simple key word concordance of a word such as *toast* into a two-dimensional scatterplot that arranges data points representing concordance lines with similar sets of context words in close spatial proximity while placing data points that have markedly different collocates further apart. Semantic patterns such as homonymy are thus reflected in different clusters of data points, yielding one cloud for tokens that signify roasted bread and

Trim: 247mm × 174mm Top: 10.5mm Gutter: 22.001mm
CUUK2979-02 CUUK2979/Kytö & Pahta ISBN: 978 1 107 03935 3 December 8, 2015 16:16



a separate one for tokens signifying that people raise a glass and drink to someone's health. In their study, Sagi et al. (2011: 171) use this procedure to investigate semantic change in the words *dog* and *deer*. The general course of the semantic developments of these elements is well-known: Old English *docga* semantically broadened so that the word *dog* today refers to not just a breed of dog, but an entire species. Conversely, Old English *deor* used to mean 'animal', today's *deer* has thus undergone semantic narrowing. Sagi et al. (2011) exhaustively retrieve examples of *dog* and *deer* from the Helsinki Corpus, construct semantic vectors for each concordance line, and visualize the results using multidimensional scaling. For the word *dog* (and its earlier spelling variants), the resulting visualization, a scatterplot of points in a two-dimensional coordinate system, reflects the process of semantic broadening. Data points from earlier corpus data occupy a smaller, more densely populated area of the scatterplot; that area grows across the subsequent corpus periods. These results align with what is generally known about the semantic development of *dog* and thus vouch for the general feasibility of the method. Beyond that, they allow a glimpse into the temporal dynamics of that development that would be hard to infer from the analysis of individual examples.

For the word *deer*, the results are less straightforward. Instead of a systematic shrinkage of the clouds of data points over time, Sagi et al. (2011: 177) observe successive shifts that show relatively little overlap between the different corpus periods. They interpret this as suggestive evidence that *deer* has undergone changes that go beyond the well-documented process of narrowing. The quantitative investigation thus prompts a more in-depth, qualitative investigation of the shifts that have taken place. What the computational procedure offers is a fresh look at data that lays bare phenomena for investigation that would have been overlooked, or perhaps considered unimportant, otherwise. Visualization techniques of the kind Sagi et al. use in that connection (see also Szmrecsanyi 2010, Hilpert 2011b) can be of considerable help for that purpose.

The rearrangement of data to facilitate qualitative analysis also lies at the heart of an exploratory analytical method that investigates shifts in the collocational behaviour of grammatical constructions (Hilpert 2006, 2008). This approach draws on the method of distinctive collexeme analysis (Gries and Stefanowitsch 2004), which is used to contrast the collocational profiles of two or more constructions that have an open slot that accommodates different lexical types. Gries and Stefanowitsch (2004: 106) exemplify the procedure with the constructions of the dative alternation. The ditransitive construction and the prepositional dative construction share a substantial number of verb types, but those shared types are not equally likely to be used in either construction. By comparing the text frequencies of both constructions against the frequencies of the verbs in either construction, verbs that significantly deviate from their expected frequencies can be identified. For instance, the ditransitive construction is significantly attracted to the verbs

Trim: 247mm × 174mm Top: 10.5mm Gutter: 22.001mm
CUUK2979-02 CUUK2979/Kytö & Pahta ISBN: 978 1 107 03935 3 December 8, 2015 16:16



*give*, *tell*, and *show* whereas the prepositional dative construction is typically used with *bring*, *play*, and *take*. These preferences are in line with the idea that the two constructions differ semantically in the distance between agent and recipient (with the ditransitive construction encoding closer proximity) and with the proposal that the ditransitive construction primarily expresses that 'X causes Y to receive Z' whereas the prepositional dative expresses that 'X causes Z to move to Y' (Goldberg 1995: 75–6). The purpose of a collostructional analysis is the exploratory semantic study of grammatical constructions via their most strongly attracted collocates. Applied to diachrony, the method can be used to contrast collocate sets of the same construction across a number of historical corpus periods. What the method provides are lists of significantly attracted collocates for each of the corpus periods that are analysed. Differences across those lists can be interpreted as a reflex of semantic change. If a construction broadens semantically, it will occur with a larger, semantically more diverse set of collocates. If a construction retreats into a particular semantic niche, it will increasingly occur with collocates that are semantically related to that niche.

The first process characterizes the development of the English *be going to* construction between the eighteenth and the twentieth century (Hilpert 2008: 120). Whereas early uses of *be going to* attract main verbs that involve animate, intentionally acting agents as their subjects, the data from later corpus periods show how the construction broadens semantically so that the attracted elements include highly general verbs such as *be* or verbs such as *happen*, which denote spontaneous events, rather than deliberate actions.

The English future construction with the modal auxiliary *shall* exhibits a very different developmental trajectory. Between the sixteenth and the twentieth century, *shall* continually decreases in text frequency and simultaneously undergoes a change towards increased usage as a text-structuring device in expressions such as *I shall return to this issue in the conclusion* or *I shall discuss quantum theory in Chapter 5* (Hilpert 2006: 252). What this suggests is that the change in question is not necessarily semantic, but rather stylistic in nature. Like the method that Sagi et al. (2011) employ to visualize phenomena of change, distinctive collexeme analysis serves to draw the analyst's attention towards those aspects of a linguistic unit that have changed over time. The quantitative method merely picks out the elements for which there is a significant difference between expected and observed frequency. A necessary second step is a qualitative analysis involving a close examination of the concrete example sentences with those significantly attracted elements, and ultimately ideally an interpretation that relates the empirical findings to a more general account of how and why the construction changed.

Other bottom-up quantitative techniques to be discussed in this overview are tools for a specific problem of diachronic corpus linguistics, namely the division of data points from different historical dates into sequential periods. All of the case studies that have been discussed up to now relied on some

Trim: 247mm × 174mm Top: 10.5mm Gutter: 22.001mm
CUUK2979-02 CUUK2979/Kytö & Pahta ISBN: 978 1 107 03935 3 December 8, 2015 16:16



contrast between earlier and later data, often with intermediate stages in between. Typically, diachronic corpus data are divided into temporal stages in a way that either captures well-established historical stages of a language or, if a more fine-grained temporal resolution is desired, in a way that uses intervals of thirty to forty years to capture changes between subsequent generations of language users. Gries and Hilpert (2008, 2012) make the case that this procedure is not without its problems. By creating equidistant time periods in a top-down kind of way, the analyst may combine corpus parts that actually behave very differently, thus creating misleading statistics/trends. Thus, one approach of Gries and Hilpert's is a data-driven approach to data periodization. The basic logic of such an approach is that (1) parts of the data that exhibit similar characteristics should form part of the same corpus period and (2) breaks between different periods should be inserted at points in time where there are measurable shifts in the characteristics of the data. Thus, periods need not be equidistant, allowing for the possibility that there are longer times of stasis that are interrupted by fits and starts of development (see Figure 2.2 for one example).

Their approach is implemented as a hierarchical clustering algorithm (Gries 2013: ch. 5). Hierarchical clustering is, like the multivariate procedures discussed earlier, a procedure that takes as its input complex datasets in which each observation (e.g., a concordance line of a particular expression in its context) is characterized in terms of a range of different variables. A common purpose of clustering approaches is to then categorize a set of $n$ observations into $m<n$ different, hierarchically ordered groups. For this purpose, the procedure compares observations, assesses their mutual similarities, and iteratively merges those two points of a set that exhibit the greatest mutual similarity. Of course, diachronic data require a small twist of that procedure. In a diachronic dataset, it is fully possible for two data points to be fairly similar despite the fact that they represent very distant historical periods. Gries and Hilpert's clustering algorithm can therefore only merge two data points if those data points are temporally directly adjacent, which motivates the algorithm's label variability-based neighbour clustering (VNC).

To illustrate the procedure and its advantages, Gries and Hilpert apply the algorithm to the data that informs Hilpert's analysis of *shall* that was discussed above. The data in question consist of collexeme strengths of the lexical verbs that occur with *shall* in corpus data from the sixteenth to the twentieth century, which combines parts of the Penn Parsed Corpora and the CLMET into a database of nearly twelve million words. The corpora from which the data is taken consist of six seventy-year periods which have been combined into pairs to yield three equidistant periods of 140 years in Hilpert's analysis. A question that VNC can answer is whether this particular periodization makes sense, or whether it would be more sensible to create different period boundaries. The raw input data for the VNC algorithm consists of six lists that contain the verbs occurring with *shall*, which add up

Trim: 247mm × 174mm Top: 10.5mm Gutter: 22.001mm
CUUK2979-02 CUUK2979/Kytö & Pahta ISBN: 978 1 107 03935 3 December 8, 2015 16:16



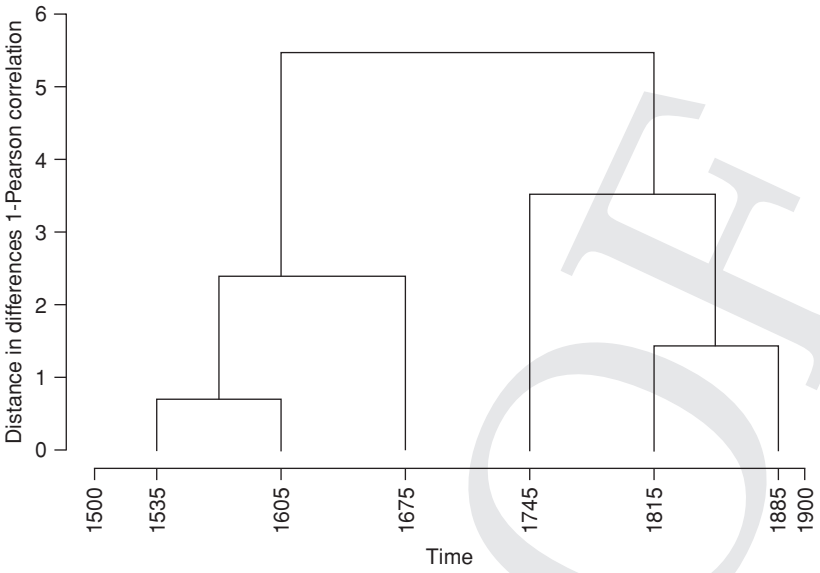


Figure 2.1 *VNC-based periodization of shall+V (Figure 3 of Gries and Hilpert 2008: 70)*

to 1,201 types, and their respective collexeme strengths. Using a correlation statistic such as Pearson's *r* (Gries 2013: sections 3.2.3 and 4.4), degrees of mutual similarity between temporally adjacent pairs can be computed. On its first iteration, the VNC algorithm finds the two lists that exhibit the greatest mutual similarity and merges the two, so that a set of only five lists remains to be analysed. The algorithm continues and finds the next pair, reiterating until all lists are merged. The result of that procedure can be visualized as a tree structure that captures mutual similarities across the six initial lists. Figure 2.1 shows the result.

The tree structure shows one thing very clearly, namely that lists three and four are not similar at all. What it suggests in terms of a reasonable periodization of the data would either be a binary split into an earlier and a later period, or a fourfold distinction of (1) periods 1 and 2, (2) period 3, (3) period 4, and (4) periods 5 and 6. A subsequent analysis that adopts such a periodization would have the benefit of seeing more pronounced differences between the corpus periods, so that statements about change can be made with greater accuracy and reliability.

A second practical benefit of VNC is that it can be used for the detection of outliers in historical data, that is, data points that deviate considerably from the overall trend in the data and/or from other temporally close data points, and that may therefore reflect 'anomalies' in the data (which in turn may result from sampling problems, author idiosyncrasies, etc.). Gries and Hilpert (2010) analyse the relative frequency of the third-person singular present tense suffixes *-(e)th* and *-(e)s* in the Parsed Corpus of Early English

Trim: 247mm × 174mm Top: 10.5mm Gutter: 22.001mm
CUUK2979-02 CUUK2979/Kytö & Pahta ISBN: 978 1 107 03935 3 December 8, 2015 16:16



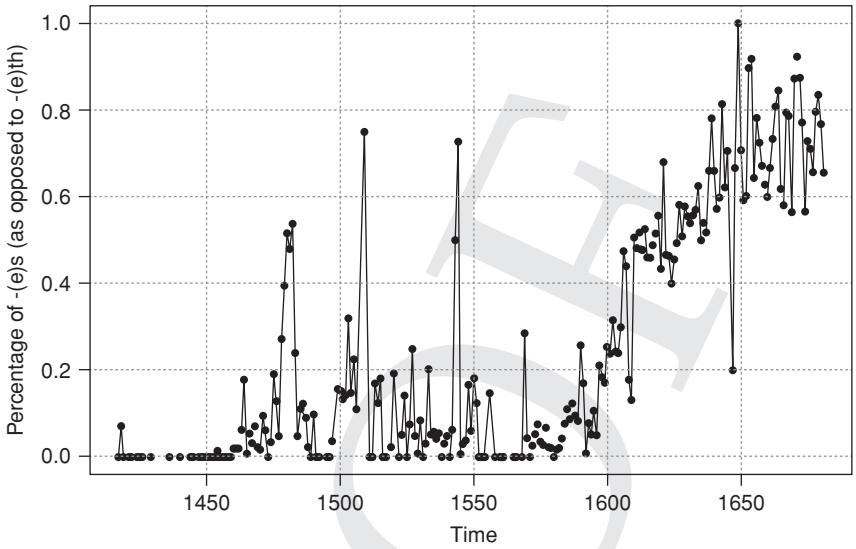


Figure 2.2 *The growth of third-person singular present tense -(e)s (from Gries and Hilpert 2010)*

Correspondence, and when the relative frequency decrease of the *-(e)th* suffix is plotted year by year, several outlier measurements are immediately apparent to the human analyst, as can be seen in Figure 2.2. Removing outliers from datasets is a practice that is common in the empirical sciences, but ideally there should be transparent conditions on outlier removal. VNC provides such conditions. Gries and Hilpert (2010: 301) use the VNC algorithm to exclude as outliers those data points that form very small clusters (i.e. an individual year) that are surrounded by much larger clusters (i.e. more than fifty years). The logic behind this approach is simple: if a measurement is a 'bad neighbour', so that it differs considerably from contemporary sources which in themselves are relatively homogeneous, this is evidence that we are dealing with an outlier.

A second type of exploratory analysis is Hilpert and Gries's (2009) iterative sequential interval estimation (ISIE). This approach is a visual tool to compare the diachronic development of observed frequencies or ratios against what would be expected on the basis of prior temporal changes and their variability. From each time period to the next, the algorithm computes and plots how more or less frequent a word/structure should be (with a kind of confidence interval) such that more recent temporal developments affect predicted developments more, and then an analyst can scan the resulting plot to determine the homogeneity of the diachronic trends and where unexpected developments occur.

In sum, exploratory tools have a lot to offer to historical corpus linguistics: they can help to discern distributional structure in data invisible to the naked eye to discover trends, temporal stages, and (un)expected diachronic

Trim: 247mm × 174mm Top: 10.5mm Gutter: 22.001mm
CUUK2979-02 CUUK2979/Kytö & Pahta ISBN: 978 1 107 03935 3 December 8, 2015 16:16



trends, which can then either inform more qualitative or additional quantitative analysis (see below for an example).

## 2.5 Desiderata for future developments

While corpus linguistics is by definition a distributional discipline and the frequencies of (co-)occurrence and dispersion statistics it provides are a natural fit with statistical methodology, the adoption of more advanced statistical tools is a slow process. Apart from very general issues that have more to do with the reporting of quantitative analyses (see Wilkinson and the Task Force on Statistical Inference 1999) that require all practitioners' attention, in this section, we will outline a few methodological approaches whose broader adoption we think would help elevate diachronic corpus linguistics to 'the next level'.

First, there are a variety of ways in which particularly variationist kinds of study can be improved. While the field is slowly discovering that generalized linear regression models are a tool much superior to traditional Varbrul analysis – recall our discussion in section 2.2 – important developments still await wider adoption. For example, simple regressions can, in fact *should*, be followed up with model criticism and evaluation to determine when levels of predictors should be conflated – does one really need to distinguish human, animate, and inanimate possessors, or is it enough to distinguish human/animate vs. inanimate? (see Bretz et al. 2010) For example, generalized linear mixed-effects models have become used more widely in linguistics as a whole because of their abilities to incorporate speaker/writer-specific effects, lexically specific effects, and to better handle unbalanced data of the kind that corpus linguists face. Thus, this method can also be very beneficial in diachronic studies. Two studies we have already mentioned showcase the power of this method: Gries and Hilpert (2010) follow up their VNC-based data periodization with such a model to study which factors drive the emergence of the *-(e)s* third-person singular, and the approach allows them to obtain a classification accuracy exceeding 94 per cent. Wolk et al. (2013) also use this method in their studies of genitives and datives, with similarly high success. While high classification accuracies are not the ultimate goal of these studies, they reflect that with the right tools, all statistics will be more precise and, thus, more relevant to the task at hand, understanding what facilitates/inhibits change. As the development of mixed-effects models (and generalized estimating equations) matures, this method will provide ever more useful results; see also Rietveld et al. (2004) for more discussion of pitfalls in quantitative corpus research.

As another example, other more sophisticated tools that can improve diachronic corpus studies involve related methods that can handle curvature/nonlinearities in the data, a frequent characteristic of diachronic data. Hilpert and Gries (2009) discuss regression with breakpoints as one rather

Trim: 247mm × 174mm Top: 10.5mm Gutter: 22.001mm
CUUK2979-02 CUUK2979/Kytö & Pahta ISBN: 978 1 107 03935 3 December 8, 2015 16:16



simple tool, but down the road more advanced methods – polynomial regressions, splines, generalized additive models (see Zuur et al. 2010 for discussion) – will also be indispensable to the quantitative study of historical corpus data.

Similar advances in the domain of exploratory tools await adoption in diachronic corpus linguistics. There is now a variety of methods to follow up on cluster-analytic results. Dendrograms such as Figure 2.1 can be studied with regard to (1) how many clusters should be distinguished (using so-called average silhouette widths), (2) how 'clean' the resulting clusters are (with F-values), (3) which features are most distinctive for the clusters (using *t*-scores), and (4) how well they map onto other cluster-analytic results regarding the same phenomenon; see Divjak and Gries (2008) for exemplification and discussion.

Finally, the statistical area of robust statistics is potentially also very useful to the study of historical corpus data. Robust statistics are statistics that are less based on the assumptions underlying many traditional statistical tools (normality, homogeneity of variances, no outliers, etc.), or that are less vulnerable in the presence of such violations, which are the rule in the kind of noisy observational data that diachronic corpus linguists study. Fields such as second-language acquisition have begun to discover this area (Larson-Hall and Herrington 2009), and in diachronic corpus linguistics, some work is under way. For instance, Lijffijt et al. (2011) develop an approach to the study of text frequency change that dispenses with the bag-of-words assumption. Much like language itself, diachronic corpus linguistics will continue to evolve.

## 2.6 Concluding remarks

This chapter has argued that quantitative methods hold considerable potential for diachronic corpus analysis. There are two main selling points. First, in order to make sense of the complex variation that is at play in processes of language change it is a simple matter of necessity to have analytical tools that can cope with that complexity and that offer the analyst a nuanced view of what happened. If we want to understand why a certain change happened, thorough understanding of how it happened is the first step towards that goal. Second, quantitative analytical methods can make phenomena visible that would otherwise not be open for inspection. These methods can offer a fresh, unbiased look at phenomena that seem familiar, but which still remain to be fully understood. Importantly, it is early days for diachronic corpus linguistics. All of the methods discussed in this chapter are still in a phase of development, awaiting further testing and replication, and we can look forward to studies that will further enlarge the toolkit of diachronic corpus linguistics in the near future. At the same time, it has to be pointed out that even a high level of analytical sophistication cannot remedy the

Trim: 247mm × 174mm Top: 10.5mm Gutter: 22.001mm
CUUK2979-02 CUUK2979/Kytö & Pahta ISBN: 978 1 107 03935 3 December 8, 2015 16:16



problem of data sparseness that is one of the natural limits of endeavour in historical linguistics. Evidently, any analytical method can only produce satisfying results on the basis of rich empirical data and analysts who know the restrictions their methods come with. When the historical record is poor, the best shot that we have at nonetheless understanding it to some degree is to take present-day variation as a model, and to see whether the historical data varies in comparable ways. To make this happen, historical corpus linguists and sociolinguistically oriented corpus linguists need to join forces both at the level of methodology and at the level of linguistic theory. Pioneering work in that direction has been carried out (Nevalainen and Raumolin-Brunberg 2003) and the recent studies that were discussed in this chapter show that the problems and pitfalls that haunt diachronic corpus linguistics are being addressed from a variety of angles.