Introduction: Investigating language variation and change

MANFRED KRUG, JULIA SCHLÜTER AND ANETTE ROSENBACH

1 The importance of variation and change

Language variation and change highlight the fact that language universally involves alternative forms and structures that compete with each other in usage. For instance, speakers of Scottish varieties of English may in certain circumstances front the initial consonant in *thing* and pronounce it as *fing*. A speaker from Cumnock in Lowland Scotland or from Portavogie in Northern Ireland may occasionally drop the subject relative pronoun in *the man (who) called me was our neighbour*. An eighteenth-century speaker and his twenty-first-century descendant may both use *kneeled down*, although the latter is more likely to use *knelt down*. As is evident from this arbitrary choice of examples from the present volume, language is inherently variable, both across time (diachronically) and at any specific point in time (synchronically). In the investigation of both synchronic and diachronic linguistic variation, the classic variables relating to the language producer are geographical, stylistic and social in nature. The fact that especially social information (like age, sex, socio-economic class) figures more prominently in the study of more recently produced data follows naturally from the fact that such information is less readily accessible for older data (cf., however, Nevalainen and Raumolin-Brunberg 2003).

In essence, the study of language variation and change investigates the ways in which language is variable, the distribution of the variants and the many factors that determine the choice of one variant over others. It has, in the past few decades, become one of the most productive and successful fields of linguistic research. Many important insights have emerged from it and enriched our understanding of the nature of language as well as how it is embedded both mentally and socially. Language variation and change is of central interest to different linguistic disciplines, such as psycho- and sociolinguistics, but also language typology, dialectology and historical linguistics. The choice of one variant rather than another more or less equivalent one tells us a great deal about the processing of the variable, about constraints on its production in real time, about characteristics of the speaker, about its social evaluation, and about general or universal tendencies in language. Furthermore, variability is a precondition for the development of diatopic and diachronic differences in language use, and its investigation provides insights into mechanisms underlying geographic distribution and historical change. Thus, variation is not only an

inherent, characteristic property of language that linguistic research has to take into account, but it is also an invaluable source of information about language itself (cf. Schlüter 2005: Chapter 1).

The present introduction situates the study of language variation and change within the discipline of linguistics and within the context of other empirical sciences (Section 2). In so doing, some fundamental concepts of empiricism are introduced (objectivity, reliability and validity; deductive vs. inductive approaches; quantitative vs. qualitative studies; Section 3). Section 4 outlines different types of empirical data and methods of data collection in linguistics and surveys the extent to which certain subdisciplines of contemporary linguistics have their 'pet' methods, i.e. affinities with certain standardized ways of collecting data. Section 5 focuses on the study of language variation and change, sketching the evolution of linguistic approaches and methods up to and focusing on the present day. We will argue in favour of a cross-fertilization between disciplines and advocate a methodological pluralism, which the present volume is designed to facilitate and promote.

2 Introspective versus empirical approaches

While the interest in language is common to all schools of linguistics, what precisely they choose as their objects of study depends significantly on their theoretical outlook. As Gilquin and Gries (2009: 1–2) remark, among linguists 'there is surprisingly little agreement on what exactly qualifies as data and how they are to be obtained, analyzed, evaluated, and interpreted'. The most striking difference in this respect is certainly the gap between purely introspective and empirically based work. To be sure, introspection can be empirically gleaned and objectivized, for instance through the use of questionnaires or experiments (see Krug and Sell, Chapter 4, or Hoffmann, Chapter 5, this volume). By 'purely introspective', however, we understand an individual researcher's intuition about a given structure (which is potentially made up during the analysis), i.e. an analysis which does not rely on a set of systematically collected data, and thus makes no obvious attempt at a supra-individual or even representative status. (Representativeness may in fact be an ultimately unattainable goal, but, in our view, one which is worth striving for.) Underlying this understanding of empirical research, then, is the concept of empiricism in the tradition of Locke and Popper, which requires systematic data collection for the formulation and falsification of hypotheses (see also Section 3 below for detail).

As is pointed out by Meyer (2009), linguistics has a long tradition as an empirical discipline. Up to the 1960s, linguists' work was based on authentic examples collected from written or spoken usage. The Chomskyan revolution and the rise of the generative paradigm put a sudden halt to this tradition, now dismissing actual language data as error-ridden and imperfect and concentrating instead on the internalized grammar of an ideal speaker/hearer. Rather than hunting for examples in the outside world, the researcher (somewhat mockingly

portrayed as an 'armchair linguist') was now justified to stay in a private study, reaching conclusions mainly based on his or her own intuition. Fillmore (1992: 36) caricatures the 'armchair linguist' as follows:

> He sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while, he opens his eyes, sits up abruptly shouting, 'Wow, what a neat fact!', grabs his pencil, and writes something down. Then he paces around for a few hours in the excitement of having come still closer to knowing what language really is like.

This approach was legitimate because generative grammarians exclusively relied on their own intuitions to determine what was grammatical or ungrammatical in their language. As Meyer (2009: 210) puts it:

> Chomsky's notion of the ideal speaker and hearer completely negated the idea of variation in language, viewing the structure of a language such as English as an idealization – a static entity not subject to any variation … For Chomsky, empirical evidence – real data and facts about usage – was less important than the particular linguistic argument being advanced.

Empirical research in current linguistics, especially from a functional perspective, is located at the other end of the methodological continuum.[1] It relies on observation, experiments and data collected from users of the language in real or purposely created situations of language use. Yet, intuition and background knowledge can and often do inspire the formulation of research questions and underlie the evaluation and classification of results.

3 Fundamental principles of empiricism

The linguistic study of variation and change – like other branches of empirical research – adheres to the three hallmarks of empiricism, namely **objectivity**, **reliability** and **validity**. The results attained should be objective, i.e. independent of the persons involved and the devices used in the study; reliable, i.e. consistent across possible repetitions of the study under identical conditions; and valid, i.e. truly pertinent to the problem under study (cf. Diekmann 2007: 250–261; Brosius, Koschel and Haas 2008: 63–70).

One thing empirical linguistics shares with more traditional generative linguistics is the link between data and theory. Depending on which of the two precedes the other in empirical research, we can distinguish between **deductive** and **inductive approaches**. On the one hand, researchers can rely on a deductive method, which is a top-down process. In the first place, principles, theories and hypotheses are formulated or adopted, and subsequently attempts are made to verify or falsify them on the basis of relevant data. This path leads from the

[1] The requirement of descriptive adequacy for natural (and variable) usage data was increasingly recognized by more recent branches of generative linguistics in the 1990s (cf. also Schlüter 2003).

general to the particular. The inductive approach, on the other hand, constitutes a bottom-up process. At first, observations and real data are analysed, which are then interpreted and used to build new hypotheses, theories or principles (which can then be confirmed or falsified in further studies). This process leads from the particular to the more general (cf. Blachowicz 2009: 310).

A further dichotomy in empirical linguistics concerns **qualitative** vs. **quantitative research**. A qualitative study typically focuses on one or a few piece(s) of evidence and analyses it in detail and with a view to a variety of its characteristics. It can investigate aspects that are difficult or impossible to quantify, such as associations, implications, opinions and feelings. In contrast, a quantitative study characteristically aims at numbers, counts or statistical measures, which is why it tends to involve a maximally large number of examples, but just a restricted set of categories for their features.

The choice of a qualitative or quantitative approach is often related to the research perspective, i.e. whether it is deductive or inductive. As Creswell (2009: 49) points out, '[i]n *quantitative research*, researchers often test theories as an explanation for answers to their questions ... In *qualitative research* ... [t]he inquirer may generate a theory as the final outcome of a study and place it at the end of a project'. Alternatively, in other qualitative studies, the theory 'comes at the beginning and provides a lens that shapes what is looked at and the questions asked'. Thus, quantitative research is usually linked to the deductive model, whereas qualitative inquiries may use inductive or deductive approaches (see Figure 1).

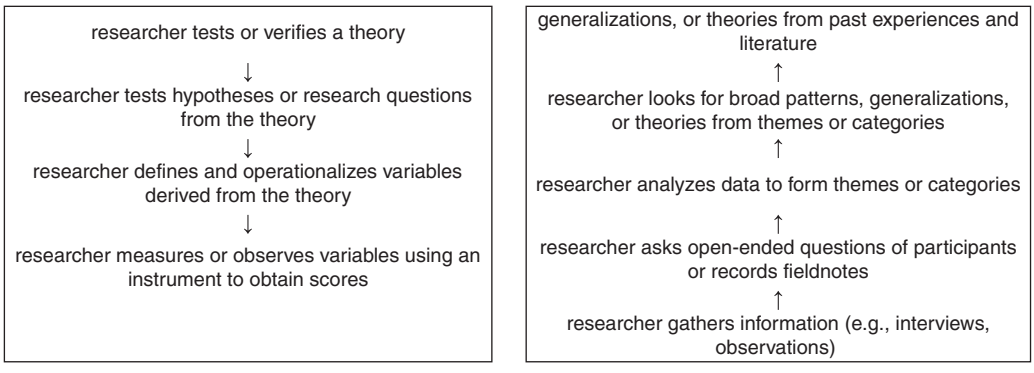

Figure 1. *The deductive approach typically used in quantitative research (left column; Creswell 2009: 57) and the inductive logic of research in a qualitative study (right column; Creswell 2009: 63)*

This dichotomy, however, does not imply that qualitative and quantitative approaches are mutually exclusive. On the contrary, a combination of methods can contribute to a better understanding of the phenomena under investigation. Mixed methods can provide rich in-depth data and ensure their generalizability to larger contexts, while at the same time avoiding the restrictions of each individual approach (cf. Angouri 2010: 33). In mixed methods research,

researchers can both test and develop theories, and they may use certain theoretical foci to guide the study.

In practice, empirical research moves forward through an alternation of induction and deduction, so that the two perspectives complement each other to ensure a steady progress. This is depicted in the so-called **empirical cycle**, which characterizes not only linguistic research, but all other empirical fields of study. Formulated in more general terms, the empirical cycle takes the following shape (cf. Figure 2).

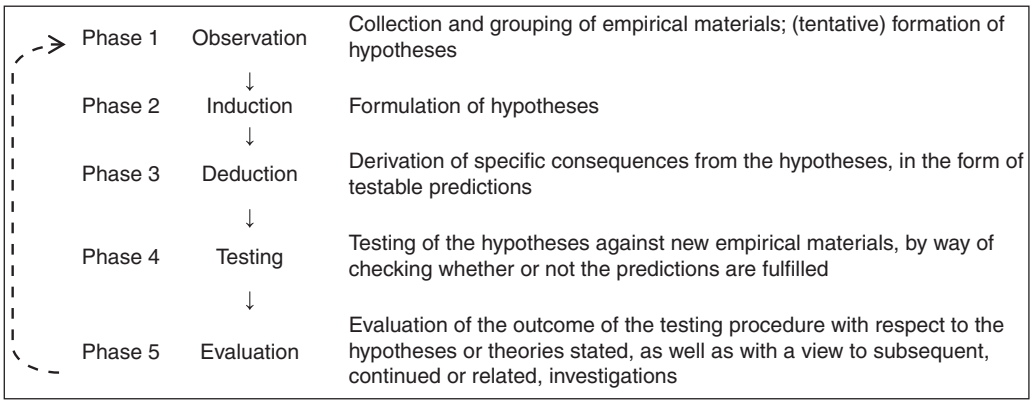

Figure 2. *The empirical cycle (adapted from de Groot's 'cycle of empirical scientific inquiry'; 1969: 28)*

Needless to say, like all models, the empirical cycle itself is an idealization as the phases do not necessarily always proceed in the chronological order given. Nevertheless, it certainly is a useful model for the design of a research project and the evaluation of previous studies.

4 Methods of data collection in empirical linguistics

Concerning the types of data used in empirical linguistics in particular, researchers have a wide spectrum of possibilities at their disposal, the choice of which depends on the specific purpose of the investigation. In principle, linguists are interested in natural, unmonitored speech production, i.e. what language users inadvertently do when not being observed. However, the presence of an observer or an experimental setting has a more or less pervasive influence on the naturalness of the communicative situation. This dilemma is known as the **observer's paradox**. There is, in general, a trade-off between the naturalness of the data and the degree of control the researcher has over them. If the researcher looks for maximally spontaneous usage produced by speakers or writers who are ignorant of the fact that their language will be used for a linguistic study, he or she has to be content with what happens to be available, for instance in a linguistic corpus or in a non-manipulated conversation. As soon

as speakers or writers are aware that their language will be investigated, e.g. in a linguistic interview situation or in front of a participant observer, they will monitor their usage to a greater or lesser extent. Elicited data, i.e. linguistic information specifically asked for by the researcher, as in a metalinguistic interview or questionnaire, is likely to be strongly influenced by the interviewee's awareness of the formality of the situation. The most artificial kind of setting in which data can be obtained is represented by invasive experiments, which may go so far as to obstruct articulatory movements. On the positive side, the more constrained the situation in which data are obtained, the better can interfering factors (or 'noise') be controlled or neutralized. The analysis can also be focused on highly specific items that may rarely be found in naturalistic data. Figure 3, which is inspired by the (partially congruent) classification provided by Gilquin and Gries (2009: 5), portrays the inverse relationship between the naturalness and the degree of monitoring inherent in linguistic data, depending on the method used for data collection.

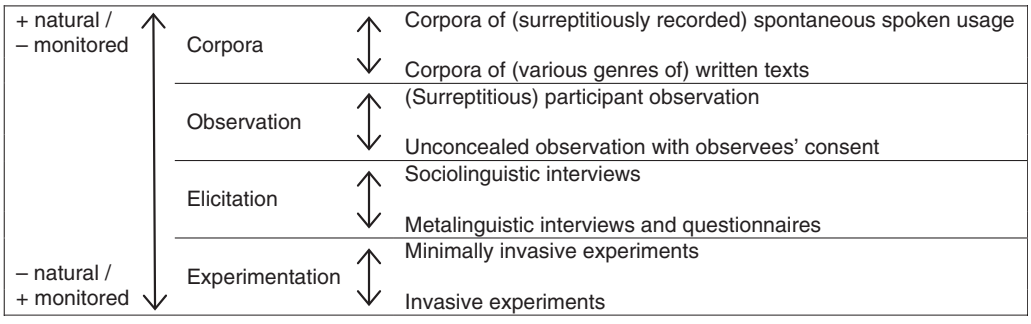

Figure 3. *Types of linguistic data, arranged according to their degree of naturalness/monitoring*

It is, thus, relatively easy to obtain a large but diffuse set of comparably spontaneous linguistic data, but it is very difficult to collect a satisfactory amount of unmanipulated data on a highly specific or infrequent linguistic phenomenon. Note, however, that Figure 3 is a gross simplification in many respects. For one thing, the actual authenticity of collected data within the categories mentioned as examples varies, in addition, with degrees of formality, register, style and editing. For instance, corpora of spoken usage, the most 'natural' data type, may be more or less monitored depending on whether they represent informal spontaneous spoken usage or planned speeches held in public. Written corpus texts may be fairly unmonitored in the case of weblogs or informal e-mail correspondence; they may, however, be extremely monitored if they are strongly edited, as in the case of newspaper texts where column width and article length matter.[2] For another thing, each method of data collection comes with its own caveats. Thus, the most promising way of obtaining

[2] Concerning the heterogeneity of spoken and written registers, see Koch and Oesterreicher (1985) or Biber (1988, 1995); see further Biber and Gray, Chapter 21, this volume.

unmonitored data, viz. the surreptitious recording of informal conversations, involves ethical if not legal problems. Note further that the most highly controlled experimental settings may unearth mere experimental effects rather than producing 'genuine' data (see e.g. Penke and Rosenbach 2004: 487–90 for further explication). (Details on the issues involved in each methodology can be found in the corresponding chapters of this volume.)

Usually, linguists from different subdisciplines have their particular 'pet' methods, which are determined by their theoretical take on language. Thus, typically, approaches focusing on the use of language prefer the most natural type of data, viz. **contemporary corpora**, as their primary data source. Sociolinguists typically elicit data on certain variables by using a technique called the **sociolinguistic interview**. Two other types of **elicitation**, namely metalinguistic interviews and questionnaires – explicitly asking informants about usage in their language or variety – are the bread-and-butter methods for typologists and dialectologists when exploring cross-linguistic and cross-dialectal similarities and differences. **Experimental data**, usually associated with the fields of psycho- and neurolinguistics, also constitute a data type specifically elicited by the researcher. Historical linguists have to cope with whatever sources are left to them. These are usually written documents, which can be compiled into **historical and diachronic corpora or databases**. This sketch is certainly grossly simplified, though it captures the broad picture of the types of methods typically applied by the different linguistic approaches.

These affinities can be documented by looking at the methods used in articles published in linguistic journals that are dedicated to individual subdisciplines. Table 1 summarizes the results of such a survey.[3] It largely confirms the picture drawn above and extends it to other subdisciplines. Those historical studies that explicitly mention an empirical method almost invariably use corpora. Research on dialects and varieties likewise depends heavily on corpus data, but also employs interviews, questionnaires and experiments. Studies in the fields of language contact, typology and anthropological linguistics nowadays incline towards corpus analysis, whereas interviews and experiments play only a secondary role. While in the aforementioned three disciplines, article abstracts more often than not leave methodological information implicit (which may be taken to indicate that they contain qualitative rather than quantitative work), the following disciplines have strong empirical foci and clearly recognizable 'pet' methods. Sociolinguistic work is very often corpus-based, but interviews are also frequently used. On the other hand, work in cognitive linguistics, psycho- and neurolinguistics, as well as in language acquisition and teaching, predominantly relies on experiments in the widest sense.

[3] Notice that the category 'corpus analysis' here embraces all methods based on a collection of naturalistic written or spoken data. For instance, it includes studies where conversations between inadvertent speakers were recorded and subsequently analysed. The category 'experiments' is another relatively broad one, covering not only highly controlled setups, but also various types of situations where participants were asked to perform a certain task.

Table 1. *Survey of methods used in different linguistic subdisciplines, according to the information given in article abstracts in recent issues of relevant journals*[(a)]

	Corpus analysis	Interviews	Questionnaires	Experiments	Other/Unclear/No indication
Historical linguistics *Diachronica* 25.1, 2008 – 26.3, 2009 *Folia Linguistica Historica* 27.1–2, 2006; 28.1–2, 2007; 29.1, 2008	7	0	0	0	40
Dialects and varieties *American Speech* 84/1–4, 2009 *English World-Wide* 29/1, 2008; 30/1–3, 2009; 31/1, 2010 *Dialectologia et Geolinguistica* 16, 2008; 17, 2009 *World Englishes* 29/1, 2010	7	2	3	3	39
Language contact, typology, anthropological linguistics *Anthropological Linguistics* 51/1, 2009 *Journal of Language Contact* 2, 2009 *Journal of Pidgin and Creole Languages* 24/1–2, 2009; 25/1, 2010 *Languages in Contrast* 9/1–2, 2009; 10/1, 2010 *Language Typology and Universals* 62/1–4, 2009	10	1	0	1	40
Sociolinguistics *Gender and Language* 3/1–2, 2009 *International Journal of the Sociology of Language* 202–203, 2010 *Journal of Sociolinguistics* 14/1–5, 2010 *Language in Society* 39/1–3, 2010	17	8	2	1	21
Cognitive linguistics, psycho- and neurolinguistics *Applied Psycholinguistics* 31/1, 2010 *Brain and Language* 113/1–2, 2010 *Cognitive Linguistics* 21/1, 2010 *International Journal of Speech-Language Pathology* 12/2, 2010 *Journal of Neurolinguistics* 23/3, 2010 *Journal of Psycholinguistic Research* 1/2010 *Language and Cognition* 1/2, 2009 *Language and Cognitive Processes* 25/3, 2010	0	2	2	30	25
First and second language acquisition and teaching *Bilingualism: Language and Cognition* 13/1–2, 2010 *English Language Teaching* 3/1, 2010 *International Journal of Bilingualism* 14/1, 2010 *Journal of Child Language* 37/1–2, 2010 *Language Learning* 60/1, 2010 *Modern Language Journal* 94/1, 2010 *Research in the Teaching of English* 44/1–2, 2009; 44/3–4, 2010 *Studies in Second Language Acquisition* 32/1, 2010	5	1	7	39	48

[(a)] This survey was conducted in collaboration with Matthias Staller, who participated in an advanced linguistics seminar taught by Julia Schlüter at the University of Regensburg in the winter term 2009/2010.

Evidently, the number of abstracts considered for Table 1 would need to be increased to gather a more representative set of data, but the results are no doubt suggestive of the most typical affinities between fields of study and methodological approaches. In sum, corpus-based research is currently the most widely spread methodology across different fields of study, particularly in those dealing with the externalized use of language, whereas experiments are the staple method of a more limited number of subdisciplines, which are characterized by a focus on mental processes in language use. What is most striking in this survey is the fact that the articles whose abstracts were surveyed hardly ever indicate that more than one methodology is applied to the phenomenon under investigation.

5 Methods in studies on language variation and change

Let us now concentrate in more detail on the field of language variation and change. Nowadays, we tend to think of linguistic variation (both synchronic and diachronic) as a research framework of its own, but essentially it is first and foremost a general linguistic phenomenon that can be, and indeed is, studied from different perspectives. Overall, the various approaches to grammatical variation follow the general tendency of individual linguistic frameworks to employ their preferred methods as outlined in the preceding section. The present book brings together some important approaches, highlighting their specific strengths and weaknesses and ways of fruitfully combining them.[4]

Dialectology is an old discipline that has always been concerned with linguistic variation, but has focused on diatopic variation (across space) rather than on the variation that can be found within individual speakers or speaker communities. The traditional methodology of dialectological studies, the drawing of dialect maps based on elicited data, encounters massive problems when intra-individual and syntopic variation comes into play (see Kretzschmar, Chapter 3, and Anderwald and Kortmann, Chapter 17, this volume).

The long-established linguistic discipline of **phonology** is not typically concerned with variation since it idealizes phonological systems on the basis of raw phonetic data. The latter do contain a substantial amount of variation, which **phoneticians** capture with a wide methodological repertoire, ranging from auditive via acoustic to articulatory techniques. These techniques are at the disposal of linguists interested in synchronic variation and its conditioning factors, though their application requires a greater or lesser amount of training (see Gut, Chapters 12 and 13, this volume).

[4] See also Rosenbach (2002: §5) for a more detailed overview of the field of grammatical variation; see Krug (2000) or Krug and Schützler (2013) for theoretically oriented approaches that, with the help of corpus linguistics and statistical analysis, integrate semantic, phonological and morphosyntactic observations into a grammaticalization framework.

While phonetic and phonological variation makes a convenient object of study in contemporary linguistics due to the high frequency of segmental and suprasegmental units, it is particularly elusive in **historical linguistics**, which has to resort to much more indirect evidence (see, e.g., Minkova, Chapter 14, this volume and the pertinent discussion in Lass 1997: Chapters 1 and 2). Yet, diachronic phonological change is a linguistic field of interest with a long tradition. In general, historical linguistics has traditionally been most concerned with variation and change in real time, for which it draws on ready-made diachronic corpora (see Chapters 6, 9 and 18 of this volume) or on collections of historical texts or quotations (see, for instance, Schlüter, Chapter 6, Rohdenburg, Chapter 7, and Minkova, Chapter 14, this volume).[5] It was Labov (1972b: Chapter 1) who also introduced apparent-time studies for analysing diachronic change. With the help of re-surveys, apparent-time studies can be replicated at a later point in time, which adds a real-time dimension to two interrelated apparent-time studies (see Fowler 1986; Bailey 2002). More recently, psycholinguistic experiments have been suggested as an additional methodological tool in historical linguistics (cf. Jäger and Rosenbach 2008).

Originally, the study of grammatical variation started within the field of **sociolinguistics**, but even here was initially limited to phonetic and phonological variation. The step from phonological to grammatical variation was a huge – and in fact rather controversial – one, as it was not clear how to define the key concept of variation studies, i.e. the linguistic variable, for the domain of morphosyntax (see e.g. Rosenbach 2002: 22–23 and references cited therein). Indeed, to this day, studies on grammatical variation from a clear sociolinguistic angle still form a minority within this framework. Sociolinguistic approaches to linguistic variation typically use spontaneous speech data, either as specifically elicited in interviews or questionnaires (see, for instance, Schreier, Chapter 1 and Krug and Sell, Chapter 4, this volume), as observed in spontaneous interaction (see Clark and Trousdale, Chapter 2, this volume), or as represented in ready-made corpora (see Mannila, Nevalainen and Raumolin-Brunberg, Chapter 18, this volume).

Emerging from the sociolinguistic approach to language variation, with its strong empirical base, has been the field of **corpus linguistics**, which is primarily defined via its preferred method, with no necessary connection to the sociolinguistic programme (see, in particular, Mair, Chapter 9, Hoffmann, Chapter 10, and Smith and Seoane, Chapter 11, this volume). The dominant data source within the field of language variation and change to this day is naturalistic spoken and written data (as represented in corpora), and accordingly the major methodological tool is corpus analysis. Some researchers have even claimed that corpus data constitute a superior data source (e.g. Leech, Francis and Xu 1994: 58, or Sampson 2001).

[5] Bauer (2002) surveys the use of corpora for tracking variation and change.