

# 3 Investigating Variation and Change in Written Documents

---

EDGAR W. SCHNEIDER

## 1 Introduction: How to Listen without Hearing

Language, Saussure taught us, is first and foremost a spoken system – writing is a secondary coding, but speech is primary (1916, repr. 1967: 45, cf. Milroy 1992: 45). In a default setting, the study of language variation and change starts out from performance data and thus employs methodological tools appropriate to the study of spoken records – sociolinguistic interviews, tape recordings, acoustic analysis, etc. However, there are areas of study for which spoken records are simply not available. In many cases we are interested in long-term developments, such as the evolution of vernaculars; and these periods of interest to linguists extend considerably beyond the time when tape recorders, or, more generally, audio recordings of speech were first available as a by-product of technological developments. It is prototypically in such instances that variation and change has to be studied on the basis of written documents only.

Normally, as variationist linguists we are not directly interested in the written record as such, not being concerned with the evolution of writing or spelling systems, questions of literacy, etc. – its function is predominantly to serve as a clue, a pathway to the variation and change of the language system in itself. More directly, most written records of interest in this context represent a speech act: either a genuine, historical one that took place at a specific time and place, recorded but indirectly in writing, or a perhaps fictional but necessarily characteristic one, rendering speech forms that a typical member of a given speech community might have uttered with some degree of likelihood, representative of the everyday communication in this community. In such cases, the written record functions as a filter, as it were: it provides us with a representation of a speech act that we would have liked to have listened to and recorded acoustically and that without the written record would have been lost altogether; but at the same time the rendering of the speech event is only indirect and imperfect, affected by the nature of the recording context in certain ways. The crucial question is to what extent the effects of the recording contexts are predictable,

or recoverable. It is essential for us to know, or to reasonably assess, what effect this filter had, how accurately the original speech event is represented. The level of accuracy may vary from a fairly faithful rendering to a gross distortion, and for the analyst it is essential to determine where on this continuum of faithfulness any given record is positioned, and what consequences this may have. In other words, whether explicitly or indirectly, a variationist linguist analyzing written records is likely to observe what I call a *Principle of Filter Removal*: a written record of a speech event stands like a filter between the words as spoken and the analyst. As the linguist is interested in the speech event itself (and, ultimately, the principles of language variation and change behind it), a primary task will be to “remove the filter” as far as possible, i.e. to assess the nature of the recording process in all possible and relevant ways and to evaluate and take into account its likely impact on the relationship between the speech event and the record, to reconstruct the speech event itself, as accurately as possible.

This chapter sets out to survey and discuss some possibilities and problems associated with this approach, and some necessary considerations and steps in the process of “filter removal”. I will begin by briefly pointing out other sub-disciplines and approaches in linguistics that have faced similar problems; obviously, variationists are neither the first nor the only group of linguists wishing to study speech through writing, and we should learn from related efforts. Subsequently, I will discuss some characteristics of the major text types available, considering their consequences for our purposes. Finally, the general methodological and theoretical problems that need to be addressed and solved as far as possible will be pointed out.

## 2 Charting the Territory: Precursors and Neighbors

Obviously, many of the questions addressed here are closely related to those historical linguists have faced for over 200 years: *historical linguistics* is also concerned with recovering earlier stages of a language (sound laws and systems, morphological categories, syntactic patterns) solely on the basis of written records that have come down to us through the centuries (cf. Cable 1990). Thus, historical linguists are also interested in collecting reliable data, evaluating sources, assessing style levels, and the role of editing (for which the philological tradition has provided principles and guidelines); in many cases they also wish to approximate speech as closely as possible (Kytö 1991: 30 lists some relevant references), and of course they face the same fundamental problem of having to cope with substantial gaps in the historical record (Thomason 1993). However, there are also differences in approaches and research interests. One concerns the application of comparative and internal reconstruction (cf. Campbell 1998: 108–48, 201–19), methods which are essential in historical linguistics but of no concern for variation studies because they assume language

uniformity – “there is nothing built into the comparative method which would allow it to address variation directly” (Campbell 1998: 146). Also, there is a difference in the time depth of investigations and in ultimate goals. In historical linguistics, the documentation of earlier stages of a language is seen as a goal in itself; traditionally, documentation and analysis have covered extended periods back to earliest records, way over a thousand years, and typically the orientation has been either strictly diachronic (i.e. describing changes of one subsystem in the course of time) or synchronic (i.e. existing forms at a given point in time). Conversely, the variationist paradigm has transcended Saussure’s claim of a separation of synchrony and diachrony (Polomé 1990: 7–8) and has typically investigated diachrony-in-synchrony, ongoing changes as reflected in social, stylistic, or linguistic distributional patterns. While variationist work can be and has been carried out as applied to early periods such as Old or Middle English, in practice the requirement of a dense documentation of vernacular speakers and styles has resulted in research activities being focused upon relatively recent periods (say, texts characteristic of nineteenth-century speech of some kind<sup>1</sup>); and while historical linguistics has had to work with poetic and formal styles to a great extent, the study of variation and change requires a focus on vernacular styles.<sup>2</sup>

*Traditional dialectology* provides us with large data sets that were mostly collected many decades ago and even then focused upon older speakers, so from a present-day perspective these records become increasingly valuable as historical data (cf. Klemola, this volume). For instance, this applies to the *English Dialect Dictionary* (Wright 1898–1906), based upon nineteenth-century collections, or to the *Linguistic Atlas of the Middle and South Atlantic States* in the USA, to which Kretzschmar and Schneider (1996: 32) explicitly ascribe a “historical orientation” and which Bailey (1997a) used in a reconstruction of variation and change in early southern English.

Another discipline that has pursued very similar goals is *historical sociolinguistics* (cf. Milroy 1992, Raumolin-Brunberg 1996, Romaine 1982), although scholars working within this framework have tended to emphasize the broader sociopolitical context of language evolution rather than detailed descriptive documentations of individual forms of a language – perhaps for want of reliable data. Milroy (1992) argues for a “variationist view of language change” (1992: 123),<sup>3</sup> and rejects tendencies to limit historical language study to the study of the history of the standard language only and to regard variability “as an obstacle rather than a resource” (1992: 132). However, although he looks into some exemplary data (for instance on *h*-dropping) of Middle English, most of the evidence in his book is derived from his modern analyses of Belfast English – essentially his proposal is programmatic only. In recent years the analysis of electronic text corpora stemming from the Helsinki school (see below) has partially redressed the balance. The most notable strictly sociolinguistic historical project to date, a remarkably successful correlation of sociohistorical class stratifications with linguistic variation, is the work done on the “Corpus of Early English Correspondence” (Nevalainen and Raumolin-Brunberg 1996).

Building a bridge between historical linguistics proper and sociolinguistics, historical *corpus linguistics* has adopted many of the incentives of the variationist paradigm (cf. Bauer, this volume). The Helsinki Corpus of historical English texts, widely used in recent years, was inspired to a considerable extent by a post-Labovian line of thinking and explicitly designed to facilitate cross-stylistic quantitative studies. Analyzing computerized samples of historical texts has turned out to be a particularly promising line of pursuit: historical records, being available in written form, lend themselves easily to computerization, and electronic corpora permit easy access to large numbers of instances of any form in context. The Helsinki school has deliberately adopted a variationist framework in their historical-linguistic analyses, as is suggested by their project title “English in transition: Change through variation,” and the fact that special attention is paid “to the role played by textual and discourse factors across the centuries” (Rissanen et al. 1997a: v, 1997b: v; cf. the papers in these volumes as well as in Hickey et al. 1997 or Nevalainen and Kahlas-Tarkka 1997) as reflected in genre and register categories. Operating within the same methodological framework, Kytö’s work on early American English (1991, on the development of modals; 1993 on third-person verbal inflection) also sees itself in a variationist tradition (1991: 83–5).

Since the 1980s the field of *pidgin and creole linguistics* has increasingly been concerned with unearthing written records as evidence of early stages of creoles. While frequently motivated by questions on the genesis of creoles, such studies have contributed significantly to our understanding of the historical evolution and the variability of these languages. Rickford (1987), for instance, uses early written texts to reconstruct the recent history of Guyanese Creole, while at the same time paying considerable attention to its internal variation. Several other creoles and pidgins, including Sranan, Negerhollands, Jamaican, Trinidadian, Bajan, Kittitian, and West African Pidgin English have received important historical documentation and analysis, mostly in book-length studies (e.g. Arends 1995, Baker and Bruyn 1999, D’Costa and Lalla 1989, Huber 1999, Rickford and Handler 1994). Theoretically and methodologically, these studies operate in a slightly different but clearly related paradigm, pursuing the same goals by means of similar approaches.

### **3 Assessing the Sources: Text Types and their Relative Proximity to Speech**

Written records that are of interest in the present context typically share certain properties and have originated in characteristic contexts. Thus, I will first consider basic requirements for texts to be acceptable for variationist analyses, then categorize them by text types, and finally discuss some characteristics of the most important text types in the light of their usefulness for variationist purposes.

### 3.1 *Some basic requirements for texts to be useful for a variationist analysis*

Only a relatively small fraction of all the historical texts that have come down to us lend themselves to a variationist analysis. Obviously, the usefulness of texts varies also individually, from one text to another, but can be broadly generalized for certain text types which share relevant discourse characteristics. Some requirements need to be fulfilled:

- 1 Texts should be as close to speech, and especially vernacular styles, as possible (Montgomery 1997a: 227). This condition largely excludes formal and literary writing – such texts may be of marginal interest, but, being shaped by prescriptive traditions and conventions, they normally display categorical, invariant usage and fail to reflect natural speech behavior and associated processes. Notably, this is at odds with the esteem attributed to texts in related disciplines; typically, we want “documents often of no particular interest to scholars in any field but linguistics” (Montgomery 1997a: 227), so there is but limited support available, and not infrequently do variationist linguists use unedited, even manuscript sources, which may cause readability problems.

**Example 1:** In compiling an electronic corpus of overseers’ letters from the pre-Civil War (“antebellum”) American South (Schneider and Montgomery, 1999), Montgomery consulted with local historians and autograph experts to eliminate undecipherable passages as far as possible.

- 2 To facilitate correlations with extralinguistic parameters, the texts should be of different origins, i.e. stem from several authors from different social classes, possibly also age groups, and both sexes, and should represent varying stylistic levels.
- 3 Texts must display variability of the phenomenon under investigation, i.e. the use of functionally equivalent variants of a linguistic variable.
- 4 With quantification being the staple methodology of variationism, texts must fulfill certain size requirements. There is no figure specifying any precise minimum number of words required – but usable texts must provide reasonably large token frequencies of individual variants, and they should (though need not) allow quantitative analyses of several phenomena, i.e. display variation in a wider range of linguistic phenomena.

### 3.2 *Categorization of text types*

While there is always some individual variation in style and expression, essentially texts come in text types, determined by their respective discourse

parameters, which, in turn, condition their proximity to speech. Thus, it will be useful to categorize text types along these lines and then to consider their individual properties in the light of their usefulness for the study of variation and change. A variety of communicative determinants of the context of situation will play a role here – whether or not texts are speech-based, whether the relationships between the participants in a discourse is close or distant, whether a communication situation is private or public, etc. (cf. Kytö 1991: 37–44). Given the requirements spelled out earlier, however, in what follows I will concentrate upon text types which bear some relatively direct relationship to speech events and ignore others, which lead us into the fuzzy boundaries between variationist and historical-linguistic analyses proper.<sup>4</sup> Adopting the Principle of Filter Removal and admitting that the “filter” may consist of a varying number of “layers,” I am proposing five text categories which represent a continuum of increasing distance between an original speech event and its written record,<sup>5</sup> based upon the following criteria:

- the reality of a speech event portrayed: a written record may be a rendition of a real and unique speech event that took place at a given time and location, or it may represent a hypothetical utterance – one that a typical member of a speech community could have made, or one that an individual would have wanted to make but was forced to make through an indirect, written channel;
- the relationship between the speaker and the person who wrote the utterance down, who may or may not have been identical; and
- the temporal distance between the speech event itself and the time of the recording (which may or may not have been simultaneous).

Table 3.1 summarizes these points and the resulting categorization. I posit the following five broad categories of the relationship between a speech event and its written record:

- 1 *Recorded*: A direct record of a singular speech event, whether written down on location and simultaneously (as in the case of trial records) or transcribed later from a mechanical recording (as in the case of Hyatt’s Hoodoo interviews).
- 2 *Recalled*: A record of a singular speech event, although written down some time after the utterance itself, presumably from notes and/or memory. The writer intends to take note of what was said verbatim and faithfully, but allowance must be made for factors such as lapses of memory or limitations of understanding. Examples: WPA ex-slave narratives, travelers’ records.
- 3 *Imagined*: A writer records potential, conceived utterances by himself which, for lack of the presence of the addressee, need to be written down rather than said; but he remains in a near-speech mode. Clearly, the boundary to genuine writing is fuzzy here, but prototypically this state of affairs characterizes writers with limited proficiency and practice in writing, who simply

**Table 3.1** Categorization of text types according to their proximity to speech

Category	Reality of speech event	speaker – writer identity	temporal distance speech – record	Characteristic text types
Recorded	real, unique	different	immediate	interview transcripts, trial records
Recalled	real, unique	different	later	ex-slave narratives
Imagined	hypothetic, unique	identical	immediate	letters, diaries
Observed	usu. real, unique	different	later	commentaries
Invented	hypothetic, unspecified	n. a.	unspecified	literary dialect

need to put their thoughts onto paper for some reason. Thus, letters by semi-literate writers belong here, but also some questionnaire responses (e.g. the *Tennessee Civil War Veterans Questionnaires*).

- 4 *Observed*: A writer cites samples of typical utterances by others that he regards as characteristic of their speech and has overheard repeatedly. Typically, such contemporaries' statements are prescriptively motivated. This is similar to category 2, except that the speech events recorded here are not unique but typical ones, and thus one step more indirect as a record of speech, filtered not only by the perception but also by the evaluation of the author.
- 5 *Invented*: This is hypothetical, imagined speech, usually thought to be uttered by others than the writer but by speakers with whose real-life models he is familiar; there is no association with a real-life speech event, but the fictitious utterance is intended to be characteristic of its – frequently also fictitious – speaker.

### 3.3 *Transcripts (Category 1: Recorded)*

Direct transcripts are clearly the most reliable and potentially the most interesting amongst all these text types – provided that they are faithful to the spoken word and the speech thus recorded represents the vernacular. Interestingly, indirect transcripts – those based upon an audio-recording of the speech event itself, can be expected to be even more accurate than direct, simultaneously

written ones, as in these cases the scribe presumably had more time to bridge the gap between speaking and writing speeds and thus to record every turn of the utterance, if desired. This is not unlike the situation of a modern sociolinguist, who has typically collected tape recordings but frequently works from transcripts of these – and the process of transcription necessarily involves some difficult decisions and some degree of subjective interpretation. Thus, transcripts of all kinds are more reliable than other types of written records, but even modern transcripts of sociolinguistic interviews are not simple representations of “reality” either, as one might think. Typically, the transcriber is the only person to have access to the audio-recording itself, so it is unusual for disagreements on transcription details to really surface in the research community<sup>6</sup> – but the problem itself clearly deserves more awareness (cf. Miethaner 2000).

It is actually not as uncommon as it may appear at first sight that a researcher has access to a written transcript of an audio-recording but not to the oral source itself. Not infrequently are transcripts of interviews but not the interviews themselves published (see example 2). In addition, there is also the more interesting (because it is diachronically relevant) case that early audio-recordings were lost but their transcripts have survived (see example 3).

**Example 2:** Loman (1967) is a collection of AAVE texts that can be read but not audited, and Patrick et al. (1996) used it as a database for a variationist analysis. Rickford (1987) contains several fieldwork transcripts, systematically put together to cover several parameters of speaker variability. In Bailey et al. (1991; cf. fn. 6) the entire transcripts of interviews with former slaves are published; the recordings themselves, however, are unpublished, available only through personal contacts.

**Example 3:** Another most promising source of earlier AAVE are the “Hoodoo”-transcripts by Harry Middleton Hyatt, used in dissertations by Ewers (1996) and – together with other sources – Kautzsch (2000a). In the 1930s, 1940s and 1970s Hyatt conducted over 1600 interviews with black practitioners of witchcraft and rootwork in 13 states. The early ones were recorded by an Ediphone with a speaking-tube, the later ones with an Ediphone and a microphone. All interviews were transcribed exactly, and the transcripts have been preserved and published, while the audio-recordings (originally on aluminium discs) were destroyed.

Direct written records of speech have been used as parts of electronic corpora and in the wake of the Helsinki school. Trial proceedings, court and meeting records, witness accounts, transcripts of sermons, and so on are “speech-based registers” (Biber and Finegan 1997: 253) which typically provide verbatim renderings of actual speech, though frequently in speech events marked by a rather formal atmosphere (Kytö 1991: 29, cf. Culpeper and Kytö 1999). Rissanen (1997) discussed the Salem witchcraft trials as linguistic evidence, in which

“particular attention was paid to every word and turn of phrase uttered by the suspects” (1997: 185).

### 3.4 *Recall protocols (Category 2: Recalled)*

These are renderings of specific speech events not taken down on the spot but at some later time from memory, possibly supported by notes. We may assume that a writer wanted to produce a transcript as faithful as possible, but perfect accuracy cannot be expected due to unavoidable distortions caused by lapses of memory and other “noise” factors, such as misperception. Psycholinguists have carried out research on what is memorized in “free recall protocols” (Hildyard and Olson 1982: 19), frequently of stories. In general, the results suggest that to some extent “surface structure features of the sentences” (1982: 19) are remembered, though a listener’s mind focuses more upon the meaning of the message than upon “the actual words, syntax and intonation,” these being rather ephemeral (1982: 20).

**Example 4:** The so-called WPA ex-slave narratives, a large-scale systematic collection of interviews with very old African Americans compiled in the 1930s (Rawick 1972/1977/1979) and analysed amongst others by Brewer (1974) and Schneider (1989), are an important source of earlier AAVE, belonging to this category. Reacting to publications which questioned their validity (see below), Schneider (1997) considered the consequences of the recording procedures in detail, and concluded that the narratives are composed of four layers of text, decreasing in their trustworthiness:

- verbatim notes;
- statements remembered accurately;
- rephrasings of the speaker’s words by the writer; and
- invented words.

Given that three of these four layers are (more or less accurate) renditions of a specific and unique speech event, and that additional evidence can be adduced for comparison and validation (like contemporary socio-linguistic analyses, or studies of other earlier records; cf. ex. 14), he believes there is no “cause for too much pessimism” (1997: 37) and classifies the narratives as “note-supported mental protocols” (1997: 44), texts composed of written notes enhanced by memory.

### 3.5 *Private letters by semi-literate writers (Category 3: Imagined)*

Clearly, letters do not represent spoken utterances; but when persons who have had but limited experience in writing and exposure to the norms of

written expression are forced to write nevertheless, their writing reflects many features of their speech fairly accurately: what they do is put their own “imagined” words onto paper, if only with difficulty. Thus, what we are most interested in are letters by semi-literate writers, a type of resource discovered, analyzed and evaluated most authoritatively in several publications by Michael Montgomery (1995, 1997a, 1997b, 1999, Montgomery et al. 1993).

**Example 5:** Montgomery (1995, 1997a, 1997b) uses Irish emigrant letters to establish transatlantic linguistic connections; he calls emigrant letters “the best resource for reconstructing early stages of American English” (1995: 5). In a very important contribution based upon letters written by African-Americans after the Civil War, Montgomery et al. (1993) document significant parallels between nineteenth-century AAVE and white dialects with respect to constraints on 3rd person plural *-s* use. Montgomery (1999) and Kautsch (2000b) use letters by African-American repatriates, from Sierra Leone and Liberia, respectively, to reconstruct features of earlier AAVE.

However useful, such letters are often products of the “vagaries and accidents of history (such as which family chose to preserve letters, whether letters survived decay)” (Montgomery 1997a: 227). One of their advantages is that they are “usually datable without ambiguity” (Montgomery 1999: 21) and “more often than not localizable to a specific place” (1999: 22; cf. Montgomery et al. 1993: 342–3). Letters “do not reflect everyday speech habits in a straightforward way [because] . . . literacy . . . always affects a person’s writing habits to some degree” (Montgomery 1995: 7), but even if they are not transcripts “with care and judgement we can separate out the evidence for speech” (1995: 5). Approximation to speech is signalled by “the lack of punctuation and other formal conventions like paragraphing” (1995: 6), unpredictable capitalization, or phonetic spellings (1995: 7).

Montgomery (1999) presents a strong principled argument in favor of the use of documents by semi-literate authors in variation studies, based on an adequate assessment of the difficulties involved. He identifies and addresses four possible problem areas:

- 1 *Authorship:* It is necessary to ask “on a case-by-case basis” (1992: 22), often in collaboration with archivists and historians, whether letters are indeed autographs (hand-written personally) or were possibly written by an amanuensis (a helper writing from dictation).
- 2 *Use of models:* The presence of opening and closing formulae and other rhetorical conventions cast doubt upon the naturalness of the speech in letters, but Montgomery argues convincingly that spelling, punctuation and other features indicate clearly that the writers do not copy from a written guide but rely on oral models, having heard letters read out aloud before. (1999: 24; cf. Montgomery 1995: 6)

- 3 *Difficulties in manipulating the written code*: While obviously even the very act of writing was difficult for many writers and many features of the letters appear “erratic and unsystematic” (1999: 24), the “conformity of many misspellings to pronunciation and the systematic patterning of grammatical features according to known constraints [show the documents to be] far from random and haphazard” (1999: 24).
- 4 *Representativeness*: Were those who could read or write not set apart from a vernacular community and its speech norms by this very ability? This is an objection which cannot be discarded but also should not be taken too seriously, Montgomery argues; in most cases the writers were not members of an elite or a distinct social group (1999: 25).

However, Kautzsch (2000a) shows that literacy results in reduced rates of vernacular forms (2000a: 189; 207) and thus keeps the Liberian letters distinct from transcript sources (which would represent categories 1 and 2 in my scheme). He suggests that certain salient nonstandard forms, including contractions, *ain't*, negative concord, zero copula (e.g. *he old*), and the nonstandard relativizers *what* and subject zero (e.g. *a man what helped me; it's the devil [Ø] makes folks do bad*), fail to surface in writing, with speakers being aware of their stigmatization (2000a: 222).

Other uses of semi-skilled letters include Bailey et al. (1989), who document the presence of the subject type constraint on plural verb-*s*<sup>7</sup> in Early Modern English, and Filppula (1999), who confirms that “private correspondence provided the most fruitful source for vernacular features” (1999: 43; cf. 43–6) in early Hiberno-English. A machine-readable corpus of antebellum overseers' letters, the compilation of which has just been completed, promises interesting insights into the nature and variability of early nineteenth-century southern US dialect (Schneider and Montgomery 1999).

A particularly interesting research strategy is suggested by Meurman-Solin (1999) in her work on early Scots. She finds an exceptionally high frequency of phonetic spellings in women's autograph letters (1999: 305), thus documenting a most interesting, culturally-based gender difference (because women had more limited access to schooling), and continues with an observation which I find remarkable: “phonetic spellings previously labeled as ‘nonstandard’ or ‘irregular’ are in fact evidence of an early adoption of later widely diffused variants. A finding of this kind may lead to a reassessment of the role of inexperienced writers as informants in the reconstruction of phonological developments” (1999: 306). This, I believe, holds great promise. It is well-known that the chronology of historical sound changes is particularly difficult to determine and tends to be fixed at the time of its completion, with little interest shown in and evidence available for the earlier stages of a change (cf. Milroy 1992: 46). If it turned out that letters allow us to detect “embryonic variants,” early traces of future changes (Gordon and Trudgill 1999), that could be an important advancement in our understanding of the mechanisms of sound change.

### 3.6 Other autograph records (Category 3: Imagined)

To the extent that semi-skilled writers wrote anything other than letters (and such writings have been preserved), such texts will be equally interesting for variationist analyses: “The unselfconscious wording scribbled down in appeals, answers and witness depositions filed to courts by untutored writers, offer unique instances of lively language . . . [and] of current colloquial usage” (Kytö 1991: 31). Such writings are rare, however, simply for lack of motivation and circumstances. Diaries may be an interesting type of text; some were used by Kytö (1991, 1993). On the other hand, the very habit of writing a diary is untypical of semi-literate writers; not surprisingly, Filppula (1999: 43) states that diaries “were disappointing in that all were written in standard language.”

There is one collection of texts that falls into this category and has been used for analyses of variation and change in early southern English, if only rarely so far: *The Tennessee Civil War Veterans Questionnaires* (Elliott and Moxley 1985). These are “first-hand reports of war-stories and attitudes of Civil War veterans” (Maynor 1993: 180), systematically collected between 1915 and 1922 in the state of Tennessee by two historians interested in writing history from below (i.e. as experienced by the common people), “a true history of the Old South” (ibid.). The questionnaire consisted of 46 questions on antebellum lifestyle, wartime experiences, etc. Some 1,650 autograph responses were submitted in response, including a wide variety of nonstandard language forms, as many of the writers were barely literate (but still willing to share their views and experiences); and they were published “exactly as written by the veterans” (ibid.). This is a very promising source, “one of very few reliable sources of data on the language of ordinary people in the nineteenth century” (Maynor 1993: 184–5), although the nature of the data, which consist largely of brief responses (except for some narrations on battles, etc.), also imposes limitations (for instance, there are many past tense verb forms, some repeating those of the questions, but no interrogatives). Little use has been made of this source so far (see Bailey 1997a and Maynor 1993 for references), but Bailey (1997a: 256) shows how data from this source can be combined with linguistic atlas records and sociolinguistic survey samples in documenting long-term change, based upon the apparent time construct.

An equally unusual but remarkable source from this category was analyzed by Bailey and Ross (1988), namely ship logs from the sixteenth through eighteenth centuries as evidence for the “Ship English” spoken by British sailors, which was the contemporary superstrate input to creolization in the New World. They faced similar problems (brevity and semantic similarity of responses; questionable representativeness because of the widespread illiteracy amongst sailors), but did find some interesting documentation of nonstandard uses, including variable constraints.

### 3.7 *Contemporary commentaries (Category 4: Observed)*

This category comprises statements on (and citations representative of) someone else's perceived, typical speech patterns, without rendering a specific speech event (in contrast with travelers' observations). Such testimony, as a manifestation of a "'negative approach' to grammar and usage [which] has a long tradition in England" (Sundby et al. 1991: 1), was typically motivated by prescriptive attitudes: observers quoted "vulgarisms" which they believed were to be avoided (but which, to turn the argument around, they perceived as being in common use around them). In historical linguistics, such evaluations have been relatively important sources of dialectal forms banned from and invisible in the written standard (cf. Tiecken-Boon van Ostade 1997). A noteworthy source, underused so far, is Sundby et al. (1991), a fairly comprehensive and systematic inventory of such statements on "vulgar" usage in England in the eighteenth century. Of course, there are also difficulties involved: such forms may have been misconceived; the representation may be distorting; usually we get isolated forms out of any context, and have little extralinguistic information on users and contexts of use – so both representativeness and validity of these examples need to be assessed with care and reluctance. Still, they do indicate earlier variation and change.

**Example 6:** Gordon (1998) provides an interesting use and evaluation of such sources in tracing the earliest stages of New Zealand English. She systematically collected comments on pronunciation in letters to newspapers and – "very valuably" (1998: 64), she says – school inspectors' reports (in which some concern on local pronunciations was voiced). Interestingly enough, the availability of archival recordings of speakers from the same period allowed her to compare and thus evaluate these comments. She finds that the written records are "reasonably reliable in certain respects" (1998: 81; e.g. with respect to *h*-dropping and the rendering of diphthongs and the centralization of /I/) but fail to record some other developments altogether, fail to comment on degrees of variability, and do not indicate earliest uses; apparently it takes a time lag for innovations to be commented on.

### 3.8 *Literary sources (Category 5: Invented)*

Literary dialect constitutes a topic in its own right – Ives (1950) is a widely-known classic, which points out its characteristics and suggests it should be assessed in comparison with modern dialect data. But it is also familiar as a source of information on variation and change (cf. several contributions in Taavitsainen et al. 1999), despite some limitations. Literary attestations have

been used “routinely, but uncritically, in attempts to document and reconstruct AA[V]E” (Montgomery 1999: 21); on the other hand, many scholars have pointed out the limitations of this approach: literary sources tend to overuse stereotypical markers but reduce variability, and they “tend to be relatively brief and open to serious questions of authenticity” (Rickford 1998: 159; cf. Maynor 1988: 110–11, Schneider 1989: 46–7). Cooley (1997) is a case in point: She shows that the speech of an African-American character in a successful eighteenth-century play is actually “based upon Caribbean varieties” (1997: 52), but nevertheless “constituted a prototype for other early African-American literary representations regardless of provenance” (1997: 53) and “became part of early American popular culture” (1997: 56). Lakoff (1982) provides a principled explanation of some of the difficulties involved. She states that the transfer of spoken discourse to fiction writing is problematic due to the different discourse requirements of the two channels (1982: 244–5) because devices of spontaneous speech function differently in written texts while, conversely, “transcripts do not feel to readers like ‘real’ conversation” (1982: 245).

Despite these reservations, literary dialect can be and has been used successfully for linguistic purposes. For instance, Mille (1997) investigates a literary representation of Gullah “as a resource to study Gullah’s history” (1997: 98), finds it helpful, and cites assessments and comparative studies that categorize it as remarkably accurate (1997: 99). Trudgill (1999) uses literary dialect to identify different degrees of salience of certain features of Norfolk dialect to Norfolk dialect writers, paying close attention to variants of nonstandard dialect orthography and their phonetic interpretation. He deduces a principle that appears generally valid for dialect orthography: “only phonological features which are currently undergoing dedialectalisation [i.e. disappearing in a dialect] are systematically represented by nonstandard dialect orthography as written by native speakers” (1999: 326).

**Example 7:** Ellis (1994) is a careful analysis of some features of Southern dialect in writings by authors from the three decades before the Civil War. Amongst other things, he documents the familiar “subject type constraint” on a broad basis, thus suggesting that “authors were using an authentic and regionally distinctive feature of early and mid-nineteenth-century Southern dialects” (1994: 135). He warns against taking literary representations at face value (1994: 128), but even after a consideration of some methodological pitfalls and limitations he argues that literary dialect should be exploited appropriately.

In using written documents for an analysis of language variation and change, it is clear that the idiosyncratic properties of each category and each individual text, resulting from its recording conditions, have to be assessed as accurately as possible to weigh its effects upon the results. In practice, however, the scarcity of useful sources recommends a broad strategy of analysing and comparing as many different sources of a single variety as possible, with results

from different text categories supplementing each other and contributing to a mutual evaluation.

**Example 8:** Though predating the variationist paradigm, Eliason (1956) is an impressive model case of an investigation of written records (in a dialectological and historical perspective). He states that after 1750 there are “plentiful” records which “reflect colloquial usage” (1956: 27) in North Carolina, and in his chapter 2 he surveys a wide range of archival manuscript sources screened for traces of vernacular language, including legal papers, bills and occupational records, plantation books and overseers’ reports, church records, children’s and students’ writings, diaries, etc.

All of the above-mentioned text categories suffer from shortcomings; and the following section will look more closely into some of the problems which they have in common. Still, despite unavoidable limitations many of the studies cited above are suggestive of the fundamental insights that can be gained from a proper analysis of such texts, and encourage researchers to go back to archives and libraries, look for appropriate texts, and investigate these in a variationist perspective – a research strategy which holds promises of substantial advances in the recovery of variation and change in earlier periods.

## 4 Problems

### 4.1 *Representativeness*

Variationists wish to understand certain principles of language organization in a speech community in general; thus, individual informants and text or tape samples are of interest only in so far as they reflect a global distribution, i.e. can be interpreted as samples drawn from and representative of a population. Representativeness is defined as the fit between a sample and the population it stands for: are we justified in assuming that the speakers and samples under investigation display the same behavior as the entire speech community? There is a crucial difference between the situations of a sociolinguist planning a present-day survey and a researcher working with historical and written data in so far as the modern sociolinguist can define a sample and select interviewees accordingly, while historical work will typically be constrained by the availability of records. Thus, assessing the representativeness of one’s sample is by necessity a crucial, unavoidable problem in working with written documents. Rather than selecting a sample from a population, the researcher typically faces the reverse situation: “The issue is not whether one has a ‘representative sample’ but to profile the sample at hand to see what inferences may reasonably be drawn from it” (Montgomery 1999: 26).

**Example 9:** In terms of sample size (1,650 respondents), the *Tennessee Civil War Veterans Questionnaires* leave nothing to be desired, and the fact that these individuals represent a broad range of social status parameters suggests there should be no problem of representativeness. However, this applies only as long as one restricts one's interests to white males, as all respondents are men, and most of them are white.

Still, many sources are voluminous enough to allow or require a selection from all the texts available. In such cases, certain simple rules should be applied to avoid any additional bias:

- Measure and constrain quality: check whether there are internal differences in validity. If so, devise a way of measuring them, and select only (from) the best sources available.
- Avoid circularity: in measuring the quality of the texts, do not use phenomena that will be the subject of the investigation proper, to avoid skewing the results.
- Diversify and stratify the sample: other considerations notwithstanding, select a diversified variety of texts (or text producers), to avoid the effects of unwanted correlations or idiolectal bias.
- All other things being equal, select a random sample: for instance, select every  $n$ -th text, with  $n$  roughly equaling the number of texts available divided by the number of texts to be sampled.

**Example 10:** In compiling a corpus of Ulster emigrant letters, Montgomery (1995) adopted only two simple selection criteria: all writers had to be from Ulster, and their letters had to attest a minimum of one nonstandard grammatical form (1995: 9). It would have been desirable to apply further criteria, but the scarcity of records did not allow any other limitations. In contrast, Schneider (1989) was lucky to be able to select a working corpus from thousands of ex-slave narratives; for these, representativeness is not a problem (though validity is). He applied a fairly elaborate procedure involving several stages and independent considerations, in line with the above recommendations (cf. 1989: 53–61).

With respect to autograph documents, there is a natural mismatch between our desire for representativeness and the quest for vernacular speech, illustrated, for example, by the fact that their sample is “skewed toward StE” in the Ship English study of Bailey and Ross (1988): most sailors were unable to write, so those who did write the ship logs were possibly not representative. This is a general problem which Montgomery identifies as a genre-specific variant of the well-known “observer’s paradox” and which he calls the “researcher’s paradox” (1997b: 125): “that individuals of lower social status whose speech intruded more directly into their writing usually wrote infrequently and were less likely to have their writing preserved” (Montgomery 1999: 26). To overcome this difficulty, Montgomery suggests an ingenious procedure, clearly related to the role of emotional questions in overcoming the observer’s paradox:

the researcher must identify persons of little education who had a compelling reason to write – preferably with some frequency, to a government official or an estate, for example – and thus who had a chance to have their letters preserved in other collections of papers. . . . there are at least three types of such individuals. These may be called lonelyhearts, desperadoes, and functionaries. If we can identify those individuals who were separated from loved ones, were in desperate straits and needed help, or were required by their occupation to submit periodic reports, we may be on the path to locating the letters of greatest interest. These three situations cut sharply across much of the social spectrum, as people of different social stations face loneliness, deprivation, or the requirement to inform others of their work, so the prospect of finding letters of less-educated persons fitting these descriptions is realistic. More important, these situations are compelling enough to motivate individuals to write for themselves, to do their best in putting words to paper regardless of their levels of literacy. In other words, someone pleading for mercy or relief may well pay little attention to the form (spelling, capitalization, grammar, etc.) of his or her writing, being more concerned with getting an unambiguous message across. The written version of the observer’s paradox is accordingly overcome – as much as is possible to do. (Montgomery 1999: 229)

It is clear that especially with small, unchangeable sample sizes representativeness can become a crucial limitation. In such cases, all that can be done is to assess the representativeness of one’s sample as well as possible, and to be reluctant in interpreting the data.

**Example 11:** Montgomery (1997b: 137) points out that the study of verbal suffixation in Early Modern English by Bailey et al. (1989) suffers from an error of representativeness: They “argue that fifteenth-century London English, represented by letters from the Cely merchant family, exhibited plural verbal –s. However, a close analysis of the data reveals that only one member of the family, Richard Cely the Younger, used the suffix regularly in nonexistential sentences; since he was reared in Yorkshire, his language most likely followed the Northern British pattern.”

Still, we should not be overly pessimistic about the limitations of written sources based on the representativeness issue – when the samples are large enough, all kinds of parameters can be investigated. Kautzsch (2000b) shows that an impressive level of sophistication can be achieved: compiling a unified corpus drawn from several independent sources, he is able to carry out apparent-time analyses of change in Earlier AAVE, with speakers’ birth years extending between the 1830s and the early twentieth century.

## 4.2 *Validity*

Validity relates to the quality of a record, its relationship to the target of investigation. If the record relates to a unique speech event (text type categories

1, 2, and 3, possibly also 4), a valid record matches what was uttered faithfully; with the other text categories, it reflects the everyday speech habits of the target community in a more global, indirect fashion. Unavoidably, the very process of writing speech down reduces validity, either because certain components of speech cannot be rendered in writing or because a writer expresses himself differently than a speaker. In autograph records, “the process of putting words into a written code operates in ways that linguists do not understand, filtering out some nonstandard forms completely, lessening the frequency of others while sometimes producing hypercorrections” (Montgomery 1999: 6). The validity of written texts for speech analyses largely depends upon the writer: his or her willingness to render speech forms, and his or her ability to do so. In addition, the recording conditions, and hence the text categories, which are shaped by these, are influential, and there is a tendency for validity to vary by text type. Thus, the validity of any individual document has to be assessed on a cline from most to least accurate. For example, in the case of travelers’ reports the validity of an observation depends on the writer’s familiarity with the variety in question, his physical proximity to the speech act itself, the temporal distance between the hearing and the taking note of an utterance, and also the discourse-pragmatic function of the written text itself. Similar considerations obtain for narratives, literary dialect, and other direct records. For example, Rissanen (1997) observes a “scale of closeness to spoken expression in the Salem documents” (1997: 185). For some text types (especially published sources), the potentially distorting influence of editing has to be considered (Maynor 1988), although that is the exception rather than the rule in the types of documents (frequently manuscripts) variationists are most interested in. Essentially, we need transcriptions which “adhere strictly to the spelling and punctuation in the original manuscript as much as possible in transcribing text to typewritten copy” (Montgomery et al. 1993: 341).

It is important and also instructive to see how an awareness of this issue has grown in the field; however, this concerns not only work with written documents but is equally valid for the assessment of other kinds of sources. Bailey (1997b) regards “evaluating data” as one of the central concerns of the discipline in the future and points out that this applies equally to written as well as spoken sources (1997b: 27–8). We are used to accepting sociolinguistic interviews and tape recordings as direct evidence, but in practice what is published and what many sociolinguists commonly work with is transcripts, and transcribing is anything but objective and unambiguous (Bailey 1997b: 28, Miethaner 2000).

**Example 12:** A growing awareness of the need to address the validity issue has become palpable in dealing with the ex-slave narratives (cf. ex. 4). Dillard (1972) was the first linguist to use selections based upon these records as illustrative examples – but he used not originals but texts from an edited selection (with linguistic modifications admitted by its editor, B. Botkin; cf. Schneider 1989: 50), and he cited these examples as representative of

current, not historical, AAE. Fasold (1976) uses the same source, Botkin's edition, as evidence for the dialect's diachrony, and briefly considers its validity (1976: 80). Brewer (1974) used the original typescripts of the narratives rather than Botkin's edition as diachronic evidence, but did not question their validity or select individual samples from the overall collection in a principled manner. Schneider (1989: 53–62) addressed these issues and thus, before selecting his sample for the analysis proper, carried out a linguistic "pre-test", a preliminary study of a feature analysed in Brewer's earlier research (copula concord, as in *I am* vs. *I is*) with the sole aim of finding out which interviewees could be assumed to have produced reliable records. Maynor (1988) compared the published versions of some of these typescripts with earlier versions in local archives and detected severe editing interference, casting doubt upon the value of these texts. In reaction, Schneider (1997) argued that the narratives combine text passages that individually can be assigned to four layers of validity (see example 4), and that these texts remain useful, within limitations (cf. Brewer 1997: 74). Kautzsch (2000a: 17–24) restricted his selection from the narrative collection to the earlier typescript versions from two states only and also regarded the interviewees as the decisive criterion for selection, admitting only texts by interviewees whose work was marked as particularly reliable by extralinguistic evidence (explicit statements on their interview practice, African-American ethnicity).

What is needed, therefore, is some means of *assessing the validity* of individual texts or collections. In what follows I propose four hierarchically ordered sets of criteria, with the higher levels indicating a higher level of validity, respectively. Each of these categories is fuzzy in itself, building upon several indicators of varying degrees of strengths; so overall a rating process will end somewhere on a continuum between relatively dubious and quite reliable validity.

- 1 *Nature of texts*: just like linguistic intuitions unavoidably influence our analyses, the surface appearance of a text, including criteria like the presence and frequency of dialectal forms, the presence of variation, and the overall impression of authenticity, plays a role in assessing a text. In practice, this is mostly a negative criterion: texts like letters, narratives, and so on will have to be excluded from further analysis if they are too close to or entirely written in the standard.
- 2 *Recording conditions*: this criterion relates to the notion of "filter removal" and the classification of text types discussed earlier: the more the recording situation is removed from the original utterance temporally, locally, and personally (from the writer's perspective), the "thicker" the filter, the less valid the rendering of text. Temporal and local distance and the degree of personal involvement are usually deducible from a text. In addition, sometimes we have explicit external evidence on the quality of a record, such as statements on interviewing and recording policy.

**Example 13:** Kautzsch (2000a) builds upon evidence of this kind when he admits texts by an interviewer whose work can be assumed to be exceptionally good, having moved into a community and acted as a participant observer. In contrast, Gordon (1998: 68) cites an interesting example of an idiosyncratic, almost haphazard limitation to the validity of early pronunciation reports in various localities across Australia and New Zealand, viz. the case of a traveler who recorded a particular pronunciation detail only after it had been explicitly pointed out to him, but subsequently he noted it in all localities – a distribution which is most unlikely to reflect real-life facts.

- 3 *Internal consistency:* in relatively large corpora, especially if they derive from several primary sources (e.g. independent sub-corpora, records by different fieldworkers, writers, or authors), it is possible to check for internal consistency. If variable features are consistently portrayed in a similar fashion, if there is a “momentum of overall consensus across fieldworkers and regions” (Schneider 1997: 37), then we can rightly assume that this reflects external reality. Similarities in constraints hierarchies and frequencies, i.e. in largely subconscious structural patterns, across independent writers and sources cannot be explained in any other reasonable way (cf. Ellis 1994: 136).
- 4 *External fit:* similarly, if the results of an investigation concur with results of other studies and familiar linguistic distributions, such as “conformity of misspellings to known phonological tendencies” (Montgomery 1999: 28), this proves the results to be trustworthy, and thus increases our overall trust in the respective source – again, it is hard to see what other cause apart from both analyses describing the same reality could explain such conformity.

**Example 14:** Maynor (1988: 115–16) admits that some of Brewer’s (1974) results match data of her own from taped interviews from the same period. In the context of the controversy over the genesis of AAVE mentioned earlier, Schneider (1983) claimed that Earlier AAVE was rich in verbal endings in all grammatical persons, a proposal which at that time was in stark contrast to conventional wisdom, which assumed an earlier creole stage of AAVE marked by a lack of inflectional endings as observed in creoles. However, his claim was supported by various observations: internal consistency of the data stemming from 104 speakers and some 40 writers (1983: 103), a regionally systematic patterning across nine states (with adjacent states showing identical inflectional systems; 1983: 105), some socially interpretable variation (1983: 106), and similarities with other descriptive statements relating to that period (1983: 103–4) as well as present-day AAVE (1983: 101) and with phonological and lexical constraints in present-day investigations (Schneider 1989: 66–71, 80–81). Schneider (1997: 41–43) shows that the written ex-slave narratives display the same rank ordering of ver-

bal –s frequencies across grammatical persons (despite significant differences in the quantities themselves) and also fairly similar frequency figures of past tense and noun plural marking as do tape recordings of socially comparable speakers. (Bailey et al. 1991)

### 4.3 *Analyzing different levels of language organization*

It is intuitively clear that the validity of speech representation will also correlate with the nature of the linguistic phenomenon under investigation: certain structural types, or, more broadly, levels of language organization are represented fairly accurately while the representation of others will be more difficult, open to doubt, or even impossible (like some phonetic details without grapheme correspondences: voicing in certain fricatives, for example). Whether or to what extent a given linguistic pattern will be represented in written encoding depends upon the following factors:

- 1 *Heaviness*: it is difficult to provide an accurate definition of the concepts of “heaviness” or “weight” (cf. Wasow 1997). For the present purpose, it is sufficient to state that heaviness correlates with length and sonority (i.e. roughly, the intensity and loudness of a given sound, measurable as its sound pressure level in air and correlating with its perceptive audibility; cf. Crystal 1991: 321–2): the longer a form is, the more phonological and morphological material is employed in its encoding, and the more sonority its phonemes have, the heavier it is, and the more likely it is to be noticed, memorized, and written down (Schneider 1997: 38). By implication, this correlates with language levels, with syntactic forms being more likely to be rendered accurately than morphological ones and, in turn, than phonological details. For example, a plural form *book-dem* is more likely to be noticed and recorded than the suffix in *books*; the same goes for a pattern like *Ain't nobody told me* as against the dental suffix in *told/tol'*, and for a past tense form *clum* as against *climbed*. Montgomery agrees that “letters offer evidence for pronunciation [but] are most amenable for the investigation of grammatical features” (1995: 7). It is worth noting that the heaviness hierarchy just sketched out applies independent of specific languages and is likely to operate universally.
- 2 *Salience*: in contrast, salience (roughly, greater awareness associated with a linguistic marker in a community, but essentially this is also a concept which is difficult to grasp; cf. Trudgill 1986) is variety-specific, but has the same effect. Features which are known to be characteristic of or socially diagnostic in a speech community are also likely to be recorded more frequently than actual usage would justify. Gordon notes this effect in observers' commentaries (1998: 68), and it is also known to occur regularly in literary dialect; in general we can expect it when a writer feels a need “to

improve the apparent authenticity of the narrative" (Schneider 1997: 38). Thus, high frequencies of occurrences of such markers in certain text types will have to be interpreted reluctantly. In contrast, inconspicuous forms, indicators which operate below the level of consciousness, are unlikely to be exaggerated or artificially inserted by a writer, and are thus more likely to be authentic when recorded. On the other hand, Kautzsch (2000a: 222) suggests that certain overtly stigmatized forms totally or largely fail to surface in writing; in his materials he finds this to be true for *ain't* (to some extent) negative concord, zero copula, and the nonstandard relativizers *what* and subject zero.

- 3 *Pragmatic and semantic conditions*: which linguistic elements are used or fail to show up in a text also depend upon its topic and pragmatic function. Most obviously this concerns the vocabulary of a text, which is frequently constrained by a narrow range of topics. However, grammatical patterns may be similarly restricted.

**Example 15:** In ship logs Bailey and Ross (1988) find "little data on indirect and direct questions, relative clauses, and modal verbs; . . . [as well as] little evidence on aspectual markers or the possible deletion of auxiliary verbs" (1988: 197–8). The extremely rare use of double modals (like *might could*) in written documents neatly illustrates the role of pragmatics: they "cluster in certain types of interactions (subtle give-and-take negotiations and sensitive face-saving situations in which highly conditional and indirect speech takes place) that are rarely found in the written record of the language" (Montgomery 1998: 96). Montgomery hypothesizes that similar restrictions might hold for double negatives, the negator *ain't* and perfective *done*. (1998: 120)

#### 4.4 *Analyzing phonetics with written records*

This is what variationists are often interested in, but what is also most difficult, given the impact of heaviness just discussed. Again, we need to find some reasonable middle ground. On the one hand, as Gordon (1998: 67–8) shows in some detail, there are some pronunciation features which orthography simply cannot render, as well as some spellings for which it is not clear what they indicate phonetically. On the other hand, "variable scribal usage is likely to be functional in some way, . . . and the most immediately obvious function of an alphabetic writing system is to relate writing to speech-forms, however complicated this relationship may be" (Milroy 1992: 142), and in a similar vein, based on his experience, Montgomery (1999: 25) confirms that "unconventional spellings almost always turn out . . . to be phonetically based in whole or in part." After all, even if the orthography of English is fairly inconsistent, there are well-established sets of grapheme-phoneme-correspondences which can be employed in writing (cf. Eliason 1956: 191–231, Miethaner 2000).

**Example 16:** Montgomery (1995: 7) discusses the phonetic interpretation of some spelling variants. He distinguishes linguistically trivial misspellings, such as a lack of double letters (*stoped*), a lack of silent letters, or spellings reflecting common pronunciation (*cuntry*, *sitty*), from meaningful spellings, such as *injoying* (suggesting the presence of the prenasalic e/i-merger) or *Prevealed* “prevealed”, *Beaker* “baker”, which implies that to the writer the spelling <ea> suggests a historically older and unraised, /e:/-like pronunciation, resulting in homophony between *reason* and *raisin*.

#### 4.5 *Choosing between qualitative and quantitative approaches*

Essentially, the variationist paradigm builds upon quantitative methodology, aiming at correlations between linguistic variants and internal or external context factors. This is also the goal of analyses of written documents, but it requires a certain breadth of coverage of variants and extralinguistic parameters of variation. Given the sampling difficulties sometimes involved in work with written documents, such frequency requirements are not always met, so quantification may not be possible or justified. Still, that does not render smaller corpora useless, but forces the analyst to resort to a more elementary level of description: if it is not possible to ask and analyse how often a variant occurs, frequently it still makes sense to ask if it occurs at all, which variants are found, and, possibly, who its users are. Two simple levels of analysis are possible underneath the level of quantification: a strictly inventorial, token-based approach (cf. Bailey and Ross 1988: 198) which limits itself to documenting which forms are found, and a slightly broader idiolect-based mode of analysis which looks both at which forms occur and which are the social characteristics of their users (ideally hoping for some pattern to emerge). A qualitative investigation is usually less sophisticated but more robust than a quantitative one, because some potentially distorting effects (such as overuse of a salient form) skew frequencies of occurrence but not necessarily the qualitative inventory of forms in a variety (cf. Schneider 1997: 43–4). It is possible that a written corpus allows reasonable (but isolated) observations, though not broader generalizations (cf. Montgomery 1995: 9).

#### 4.6 *Determining extralinguistic context parameters*

Finally, in working with semi-vernacular texts it may be difficult to adequately categorize and manipulate the extralinguistic context parameters of interest, especially style and class. Frequently an assessment of these parameters requires culture-specific knowledge that may be hard to obtain. Familiar categories derived from present-day investigations may prove of little value in earlier times and other contexts. For example, Nevalainen (1996: 58–61) discusses the

social hierarchy in Tudor England with the aim of establishing a model of “social class” distinctions appropriate for early England, and she arrives at a class stratification that looks alien to a modern sociolinguist, comprising strata like “nobility”, “lower gentry”, “upper clergy”, “merchants”, etc. In some cases very little may be known at all: according to Montgomery (1999: 21), “lack of personal information about the writers” is characteristic in the case of documents by semi-literate writers”, because “less-educated people didn’t usually write to strangers” (1997a: 228). Estimating the stylistic level of some documents may be equally difficult. The only solution to this problem is to collect as much relevant factual information as possible from expert sources, to arrive at a maximally appropriate ranking of styles, status levels, and the like.

## 5 Conclusion: Pitfalls and Advantages

As is conveniently summarized by Montgomery et al. (1993: 345), analyses of written documents hold specific “potential pitfalls for linguists. Assessing the degree of their vernacularity is a crucial issue that must always be addressed, though it is very often slighted by linguists. Written documents inevitably conceal some, perhaps many, of the speech patterns of their authors and can never be taken at face value as the equivalent of transcripts, especially for phonological purposes.” It is necessary to assess the characteristics of text types and individual texts in the light of their historical and culture-specific settings; it is necessary to judge the representativeness of one’s sample as well as the validity of a group of texts or a single source, and it is necessary to consider the possible effects of these factors upon the representation of a given linguistic level or feature with care, judgment, and reluctance: it is to be expected that these vary greatly from one source to another, from one goal of analysis to another. Mostly the nature of the sources available will determine, sometimes limit, what can be achieved with them.

But that is not to say that analysing written sources is a second-best solution by necessity. It is important to understand that essentially the same considerations and sometimes also limitations hold, *mutatis mutandis*, for tape-recorded surveys as well, so perhaps the above considerations also serve to sharpen our eye for essentially similar requirements in working with all kinds of real-life language, a need for the qualification, assessment and interpretation of one’s sources (cf. Bailey 1997b: 27–8). Perhaps written documents are one further step removed from the “reality” of speech performance than tape recordings. However, tape recordings are also not the direct road to “truth” without further considerations: transcriptions are anything but objective, as Miethaner (2000) shows convincingly, and the need to deal with parameters of cultural context, class, and style (including, for instance, the observer’s paradox), arises in much the same way. Working with written data requires somewhat more judgment and assessment than an analysis of audio recordings, but the difference is a

matter of degree: essentially, with both approaches the goal is the same, and the pathways to reach it are very similar.

## NOTES

- 1 A case in point is the history of African American Vernacular English (AAVE): this dialect has been of particular importance to variationists because of the sociopolitical and linguistic interest in it and because its origins have been contested, with proposals ranging between a largely creole-derived and African-influenced genesis and a predominantly British-dialectal and archaic character. Thus, many of the examples to be discussed below reflect efforts to uncover the early history or the nineteenth-century character of AAVE.
- 2 Obviously, the boundary between the two approaches is notoriously and unavoidably fuzzy. There is some degree of overlap in the fundamental interest in documenting systematic patterns of change, in the search for precise conditions of and constraints on the use of functionally similar alternatives, and in an emphasis on taking various style levels into account – studies marked by these three features are clearly of interest in both paradigms. As an example, take Ball (1994), an investigation of *it*-clefting based upon “treatises, essays and fiction, representing, it was hoped, a continuum from learned to more colloquial prose” (1994: 184; cf. Ball 1996 on relative pronoun choice). An emphasis on ordered variability with a strong theoretical interest and the uses of quantitative methods are relatively more characteristic of the “variation and change” paradigm than of a traditional historical-linguistic framework. For thoughts on the distinction between “traditional” and “variational” historical approaches, see also Kytö (1991, e.g. 70–1).
- 3 With respect to the study of Middle English, he states: “One of the advantages of studying Middle English is that its written forms are highly variable. . . . not only is there considerable divergence *between* different texts, there is also normally great variability (particularly in spelling and inflexional forms) *within* the texts. Thus, ME language states, being so variable, should in principle be suited to the same kind of analysis that we use in present-day social dialectology, and by using variationist methods we should be able to explore at least some of the constraints on variation that might have existed in ME. . . . in ME we must locate these constraints initially through the writing system” (Milroy 1992: 131).
- 4 Such marginally relevant text types comprise what Kytö (1993: 117) calls “scripts”, “texts written to be spoken”, including sermons, for instance.
- 5 A comparable categorization of early dialogue texts is proposed by Culpeper and Kytö (1999), who distinguish the following three types: “(1) *Recorded* – texts produced from notes taken down by an individual, such as a clerk, present at a particular speech event; (2) *Reconstructed* – texts which purport to present dialogue which actually took place at some point in the past (usually, the

- narrator was present at the speech event in question); and (3) *Constructed* – texts which contain constructed imaginary dialogue” (1999: 173).
- 6 A case in point is the variant readings of the ex-slave tapes whose transcripts are published and interpreted in Bailey et al. (1991), a well-known and widely discussed resource for earlier AAVE. The authors explicitly describe the process of composing the transcripts, which went through five separate audittings and still left some points of disagreement marked (1991: 14–17). Nevertheless, Rickford (1991) challenged some interpretations of these recordings, claiming that the transcript is unreliable and potentially biased. Most recently, Sutcliffe (2001) has built far-reaching assumptions on the origin of AAVE on the presumed presence of weak, admittedly doubtful and almost inaudible forms on these tapes that everybody but he, he claims, has failed to hear so far.
- 7 This is a tendency, documented in some dialects of English, for a third person plural predicate verb to be marked with a suffix *-s* after a full noun phrase subject but to have no such suffix after a pronoun subject, i.e. a preference for forms like *the dogs barks* but *they bark*.

## REFERENCES

- Arends, Jacques (ed.) (1995). *The Early Stages of Creolization*. Amsterdam, PA: John Benjamins.
- Bailey, Guy (1997a). When did Southern English begin? In Edgar W. Schneider (ed.), *Englishes around the World 1: General Studies, British Isles, North America*. Amsterdam: John Benjamins. 255–75.
- Bailey, Guy (1997b). Southern American English: a prospective. In C. Bernstein, Th. Nunnally and R. Sabino (eds.), *Language Variety in the South Revisited*. Tuscaloosa and London: University of Alabama Press. 21–31.
- Bailey, Guy, Natalie Maynor, and Patricia Cukor-Avila (1989). Variation in subject-verb concord in Early Modern English. *Language Variation and Change* 1: 285–300.
- Bailey, Guy, Natalie Maynor and Patricia Cukor-Avila (1991). *The Emergence of Black English: Text and Commentary*. Amsterdam: John Benjamins.
- Bailey, Guy and Garry Ross (1988). The shape of the superstrate: Morphosyntactic features of Ship English. *English World-Wide* 9: 193–212.
- Baker, Philip and Adrienne Bruyn (eds.) (1999). *St. Kitts and the Atlantic Creoles: The Texts of Samuel Augustus Mathews in Perspective*. London: University of Westminster Press.
- Ball, Catherine N. (1994). Relative pronouns in it-clefts: The last seven centuries. *Language Variation and Change* 6: 179–200.
- Ball, Catherine N. (1996). A diachronic study of relative markers in spoken and written English. *Language Variation and Change* 8: 227–58.
- Bernstein, Cynthia, Thomas Nunnally and Robin Sabino (eds.) (1997). *Language Variety in the South Revisited*. Tuscaloosa and London: University of Alabama Press.
- Biber, Douglas and Edward Finegan (1997). Diachronic relations among

- speech-based and written registers in English. In T. Nevalainen and L. Kahlas-Tarkka (eds.), *To Explain the Present. Studies in the Changing English Language in Honour of Matti Rissanen*. Helsinki: Société Néophilologique. 253–76.
- Brewer, Jeutonne P. (1974). The verb *be* in Early Black English: A study based on the WPA ex-slave narratives. Ph.D. dissertation, University of North Carolina at Chapel Hill.
- Brewer, Jeutonne P. (1997). Challenges and problems of recorded interviews. In C. Bernstein, Th. Nunnally and R. Sabino (eds.), *Language Variety in the South Revisited*. Tuscaloosa: University of Alabama Press. 59–75.
- Cable, Thomas (1990). Philology: Analysis of written records. In E. Polomé (ed.), *Research Guide in Language Change*. Berlin and New York: Mouton de Gruyter. 97–106.
- Campbell, Lyle (1998). *Historical Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.
- Cooley, Marianne (1997). An early representation of African-American English. In C. Bernstein, Th. Nunnally and R. Sabino (eds.), *Language Variety in the South Revisited*. Tuscaloosa and London: University of Alabama Press. 51–8.
- Crystal, David (1991). *A Dictionary of Linguistics and Phonetics*. 3rd edn. Oxford: Blackwell.
- Culpeper, Jonathan and Merja Kytö (1999). Investigating nonstandard language in a corpus of Early Modern English dialogues: Methodological considerations and problems. In I. Taavitsainen, G. Melchers and P. Pahta (eds.), *Writing in Nonstandard English*. Amsterdam: John Benjamins. 171–87.
- D'Costa, Jean and Barbara Lalla (eds.) (1989). *Voices in Exile: Jamaican Texts of the 18th and 19th Centuries*. Tuscaloosa: University of Alabama Press.
- Dillard, Joey L. (1972). *Black English: Its History and Usage in the United States*. New York: Random House.
- Eliason, Norman E. (1956). *Tarheel Talk: An Historical Study of the English Language in North Carolina to 1860*. Chapel Hill: University of North Carolina Press.
- Ellis, Michael (1994). Literary dialect as linguistic evidence: subject–verb concord in nineteenth-century southern literature. *American Speech* 69: 128–44.
- Ewers, Traute (1996). *The Origin of American Black English: Be-forms in the HOODOO texts*. Berlin and New York: Mouton de Gruyter.
- Fasold, Ralph (1976). One hundred years from syntax to phonology. In Sanford B. Steever, Carol A. Walker and Salikoko S. Mufwene (eds.), *Papers from the Parasession on Diachronic Syntax*. Chicago: Chicago Linguistic Society. 79–87.
- Filppula, Markku (1999). *The Grammar of Irish English: Language in Hibernian Style*. London: Routledge.
- Gordon, Elizabeth (1998). The origins of New Zealand speech: The limits of recovering historical information from written records. *English World-Wide* 19: 61–85.
- Gordon, Elizabeth and Peter Trudgill (1999). Shades of things to come: Embryonic variants in New Zealand English sound changes. *English World-Wide* 20: 111–24.
- Hickey, Raymond, Merja Kytö, Ian Lancashire and Matti Rissanen (eds.) (1997). *Tracing the Trail of Time. Proceedings from the Second Diachronic Corpora Workshop*. Amsterdam and Atlanta: Rodopi.

- Hildyard, Angela and David R. Olson (1982). On the comprehension of oral vs. written discourse. In D. Tannen (ed.), *Spoken and Written Language: Exploring Orality and Literacy*. Norwood, NJ: Ablex. 19–33.
- Huber, Magnus (1999). *Ghanaian Pidgin English in its West African Context. A Sociohistorical and Structural Analysis*. Amsterdam and Philadelphia: John Benjamins.
- Ives, Sumner (1950). A theory of literary dialect. *Tulane Studies in English* 2: 137–82.
- Kautzsch, Alexander (2000a). The historical evolution of earlier African American English: A comparison of written sources. Ph.D. dissertation, University of Regensburg.
- Kautzsch, Alexander (2000b). Liberian letters and Virginian narratives: Negation patterns in two new sources of earlier AAE. *American Speech* 75: 34–53.
- Kretzschmar, William A. Jr and Edgar W. Schneider (1996). *Introduction to Quantitative Analysis of Linguistic Survey Data: An Atlas by the Numbers*. Thousand Oaks, CA: Sage.
- Kytö, Merja (1991). *Variation and Diachrony, with Early American English in Focus*. Frankfurt: Peter Lang.
- Kytö, Merja (1993). Third-person present singular verb inflection in early British and American English. *Language Variation and Change* 5: 113–39.
- Lakoff, Robin Tollmach (1982). Some of my favorite writers are literate: The mingling of oral and literate strategies in written communication. In D. Tannen (ed.), *Spoken and Written Language: Exploring Orality and Literacy*. Norwood, NJ: Ablex. 239–60.
- Loman, Bengt (1967). *Conversations in a Negro American Dialect*. Washington, DC: Center for Applied Linguistics.
- Maynor, Natalie (1988). Written records of spoken language: How reliable are they? In A. R. Thomas (ed.), *Methods in Dialectology*. Clevedon, PA: Multilingual Matters. 109–20.
- Maynor, Natalie (1993). Reconstructing nineteenth-century Southern White English: more evidence from *The Tennessee Civil War Veterans' Questionnaires*. In W. Viereck (ed.), *Historical Dialectology and Linguistic Change: Proceedings of the International Congress of Dialectologists, Bamberg*, vol. 2. Stuttgart: Steiner. 180–90.
- Meurman-Solin, Anneli (1999). Letters as a source of data for reconstructing early spoken Scots. In I. Taavatsainen, G. Melchers and P. Pahta (eds.), *Writing in Nonstandard English*. Amsterdam: John Benjamins. 305–22.
- Miethaner, Ulrich (2000). Orthographic transcriptions of nonstandard varieties: the case of earlier African American English. *Journal of Sociolinguistics* 4: 534–60.
- Mille, Katherine Wyly (1997). Ambrose Gonzales's Gullah: What it may tell us about variation. In C. Bernstein, Th. Nunnally and R. Sabino (eds.), *Language Variety in the South Revisited*. Tuscaloosa: University of Alabama Press. 98–112.
- Milroy, James (1992). *Linguistic Variation and Change: On the Historical Sociolinguistics of English*. Oxford and Cambridge, MA: Blackwell.
- Montgomery, Michael (1995). The linguistic value of Ulster emigrant letters. *Ulster Folklife* 41: 1–16.
- Montgomery, Michael (1997a). A tale of two Georges: The language of Irish Indian traders in colonial North America. In Jeffrey Kallen (ed.), *Focus on Ireland*. Amsterdam: John Benjamins. 227–54.
- Montgomery, Michael (1997b). Making transatlantic connections between

- varieties of English: The case of plural verbal –s. *Journal of English Linguistics* 25: 122–41.
- Montgomery, Michael (1998). Multiple modals in LAGS and LAMSAS. In M. Montgomery and Th. Nunnally (eds.), *From the Gulf States and Beyond: The Legacy of Lee Pederson and LAGS*. Tuscaloosa: University of Alabama Press. 90–122.
- Montgomery, Michael (1999). Eighteenth-century Sierra Leone English: Another exported variety of African American English. *English World-Wide* 20: 1–34.
- Montgomery, Michael, Janet M. Fuller and Sharon DeMarse (1993). “The black men has wives and Sweet harts [and third person plural –s] Jest like the white men”: Evidence for verbal –s from written documents on 19th-century African American speech. *Language Variation and Change* 5: 335–57.
- Nevalainen, Terttu and Leena Kahlas-Tarkka (eds.) (1997). *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*. Helsinki: Société Néophilologique.
- Nevalainen, Terttu and Helena Raumolin-Brunberg (eds.) (1996). *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence*. Amsterdam: Rodopi.
- Patrick, Peter et al. (1996). 100 years of (TD)-deletion in African American English. Paper given at NWAVE25, Las Vegas.
- Polomé, Edgar C. (1990). Language change and the Saussurean dichotomy: Diachrony versus synchrony. In E. Polomé (ed.), *Research Guide in Language Change*. Berlin and New York: Mouton de Gruyter. 3–9.
- Raumolin-Brunberg, Helena (1996). Historical sociolinguistics. In T. Nevalainen and H. Raumolin-Brunberg (eds.), *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence*. Amsterdam: Rodopi. 11–37.
- Rawick, George P. (ed.) (1972/1977/1979). *The American Slave: A Composite Autobiography*, 19 vols.; Supplement Series 1, 12 vols.; Supplement Series 2, 10 vols. Westport, CT: Greenwood.
- Rickford, John R. (1987). *Dimensions of a Creole Continuum: History, Texts & Linguistic Analysis of Guyanese Creole*. Stanford, CA: Stanford University Press.
- Rickford, John R. (1991). Representativeness and reliability of the ex-slave materials, with special reference to Wallace Quarterman’s recording and transcript. In G. Bailey, N. Maynor and P. Cukor-Avila (eds.), *The Emergence of Black English: Text and Commentary*. Amsterdam: Benjamins. 191–212.
- Rickford, John R. (1998). The creole origins of African-American Vernacular English: evidence from copula absence. In Salikoko S. Mufwene, John R. Rickford, Guy Bailey and John Baugh (eds.), *African-American English: Structure, History and Use*. London: Routledge. 154–200.
- Rickford, John R. and Jerome Handler (1994). Textual evidence on the nature of early Barbadian speech, 1676–1835. *Journal of Pidgin and Creole Languages* 9: 221–55.
- Rissanen, Matti (1997). “Candy no witch, Barbados”: Salem witchcraft trials as evidence of Early American English. In Heinrich Ramisch and Kenneth Wynne (eds.), *Language in Time and Space*. Stuttgart: Steiner. 183–93.
- Rissanen, Matti, Merja Kytö, and Kirsi Heikkonen (eds.) (1997a). *English in Transition: Corpus-Based Studies in Linguistic Variation and Genre Styles*.

- Berlin and New York: Mouton de Gruyter.
- Rissanen, Matti, Merja Kytö and Kirsi Heikkonen (eds.) (1997b). *Grammaticalization at Work: Studies of Long-Term Developments in English*. Berlin and New York: Mouton de Gruyter.
- Saussure, Ferdinand de. (1916, repr. 1967). *Cours de linguistique générale*. Paris: Payot.
- Schneider, Edgar W. (1983). The origin of the verbal –s in Black English. *American Speech* 58: 99–113.
- Schneider, Edgar W. (1989). *American Earlier Black English: Morphological and syntactic variables*. Tuscaloosa: University of Alabama Press.
- Schneider, Edgar W. (1997). Earlier Black English revisited. In C. Bernstein, Th. Nunnally and R. Sabino (eds.), *Language Variety in the South Revisited*. Tuscaloosa: University of Alabama Press. 35–50.
- Schneider, Edgar W. and Michael Montgomery (1999). On the trail of early nonstandard grammar: An electronic corpus of Southern US antebellum overseers' letters. Paper given at NWAVE28, Toronto.
- Sundby, Bertil, Anne Kari Bjørge and Kari E. Haugland (1991). *A Dictionary of English Normative Grammar, 1700–1800*. Amsterdam: Benjamins.
- Sutcliffe, David (2001). The voice of the ancestors: new evidence on 19th-century precursors to 20th-century African American English. In S. Lanehart (ed.), *Sociohistorical and Cultural Contexts of African-American English*. Amsterdam: Benjamins.
- Taavitsainen, Irma, Gunnel Melchers and Päivi Pahta (eds.) (1999). *Writing in Nonstandard English*. Amsterdam: Benjamins.
- Thomason, Sarah Grey (1993). Coping with partial information in historical linguistics. In H. Aertsen and R. J. Jeffers (eds.), *Historical Linguistics 1989*. Amsterdam: Benjamins. 485–96.
- Tiecken-Boon van Ostade, Ingrid (1997). Lowth's corpus of prescriptivism. In T. Nevalainen and L. Kahlas-Tarkka (eds.), *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*. Helsinki: Société Néophilologique. 451–63.
- Trudgill, Peter (1986). *Dialects in Contact*. Oxford and New York: Blackwell.
- Trudgill, Peter (1999). Dedialectalisation and Norfolk dialect orthography. In I. Taavitsainen, G. Melchers and P. Pahta (eds.), *Writing in Nonstandard English*. Amsterdam: Benjamins. 323–9.
- Wasow, Thomas (1997). Remarks on grammatical weight. *Language Variation and Change* 9: 81–105.
- Winter, Werner (1990). Linguistic reconstruction: The scope of historical and comparative linguistics. In E. Polomé (ed.), *Research Guide in Language Change*. Berlin and New York: Mouton de Gruyter. 11–21.
- Wright, Joseph (1898–1906). *The English Dialect Dictionary*, 6 vols. London: Oxford University Press.