Using historical texts

Chapter 11:
Using historical texts

Ans van Kemenade
Bettelou Los

1. Introduction

> The fact is, Phaedrus, that writing involves a similar disadvantage to painting. The productions of painting look like living beings, but if you ask them a question, they maintain a solemn silence. (Plato, *Phaedrus*, tr. Walter Hamilton 1975: 96)

This chapter is on how we can bring the evidence from textual data to bear on linguistic analysis, primarily in historical linguistics. Linguistic analysis comes in many varieties, however, and before we discuss the value of textual data, we should briefly consider what the object of study of linguistics is. Much depends here on theoretical perspective: to mention some examples, linguists concerned primarily with language use may be interested in spoken language use, variation in register, interactive modes, language use as a marker of social status, including prestige-driven norms, and so on. Linguists working from a formal perspective will be interested in speakers' language competence, the internalized grammar that is assumed to be the core of a speakers' knowledge of language.

These various types of linguists have very different objects of study, but with respect to the use of textual data, they have an important thing in common: written language is a derivative, situated at some remove from the chosen object of linguistic investigation. This position has long been recognized: Delbrück states in his *Introduction to the Study of Language*, a Neogrammarian "manifesto," that "The guiding principles for linguistic research should accordingly be deduced not from obsolete written languages of antiquity, but chiefly from the living popular dialects of the present day"(Delbrück 1882: 61). De Saussure, too, notes that "writing is foreign to the internal system of the language. Writing obscures our view of the language, writing is not a garment, but a disguise" (de Saussure 1983: 24). Linguistic research, however, often has to rely on written texts; the linguist interested in the syntax or lexicon in language use, unlike for instance phonology, requires a very large database in order to ensure that there is a reasonable chance that it contains a more or less full range of constructions or lexical items, and the collection (recording and transcription) of spontaneous speech is more costly than the collection of written texts. Another motivation for studying written texts is to study the effect of written conventions on the spoken language: in languages that are highly standardized, the prescriptive norms of the standard language are likely to influence the spoken language use of speakers, either consciously or unconsciously. Linguists may, of course, also be interested in the language of texts in its own right, for the study of genres and writing conventions, for stylistic purposes, for the analysis of language ideologies (see Chapter 21), or for the study of narrative and other literary techniques (see Traugott and Pratt 1980).

This chapter is aimed at readers who plan to use texts for the purpose of linguistic analysis. Many of the methodological points made in this chapter are likely to be equally valid for other types of textual research, as these must base themselves on an interpretation of the

language evidence as well. The chapter will focus on the use of textual data in historical linguistics, a field that cannot employ data collection methods typically used with native speakers, such as introspection (Chapter 3), elicitation (Chapter 4), questionnaires (see Chapter 6), or experiments (Chapter 7). While for the study of present-day language use, linguists have access to sources for spoken and written language use, historical linguistic research must by its very nature base itself on written texts. We will discuss some pitfalls and caveats that follow from this in section 2. Section 3 will discuss the use of textual material for the sociolinguistic study of language change. Section 4 will focus on electronic text corpora, which has made data gathering much quicker, and allows us to resolve some of the pitfalls discussed in sections 2 and 3.

Other caveats and pitfalls have to do with misinterpretations or misrepresentations of the data, particularly when investigators rely on databases that were created by others in order to answer a specific research question, which may mean that crucial context is missing. Most of these problems boil down to a failure to compare like with like: drawing conclusions from samples that not only differ in historical period but also in dialect, or in genre or register, or in data type. These caveats will be discussed in section 5. There are ways of creatively making "the best use of bad data" (Labov 1994:11); we will discuss some examples of circumventing data gaps in section 6.

Throughout the chapter, our examples will be drawn primarily though not exclusively from historical English. They will , however, be framed in such a way as to bring out their general relevance.

2. Studying language change through written texts

Linguists using texts to study language structure must infer properties of historical stages of spoken language from written evidence. However, oral and written language can diverge in a number of ways. Authors of written texts, unlike speakers in natural conversational settings, cannot rely on immediate hearer feedback to repair hitches in communication, but have to anticipate such hitches by being more explicit and expressive than they would have been as speakers. When speakers become authors and hearers readers, they cannot rely on cues from prosody and intonation but have to find different ways of getting their message across. Written styles accordingly differ from oral styles by their use of compensatory strategies to help the reader through the text. Such styles do not develop overnight, but require a literary culture, which in turn depends for its development on rates of literacy and the availability of texts. Studies of oral versus literate strategies suggest that in literate traditions "the meaning is in the text," in the actual written words, while in oral situations "the meaning is in the context" and in the implications of communicative acts (Fleischman 1990: 22, quoting Goody & Watt 1968; see also Olson 1977; Bauman 1986). Texts from earlier periods often reflect oral speech styles more closely: they use parataxis (strings of loosely connected main clauses) rather than hypotaxis (subclauses embedded in main clauses), and discourse particles whose functions are difficult to identify, repetitions, unexpected resumptive pronouns, left dislocations, and inconsistent use of tenses (Fleischman 1990: 23).

Literate traditions develop stylistic conventions in writing (Perret 1988). Other conventions develop as the result of explicitly formulated views. Lenker's (2010) study charts the development of new written styles once English, in the course of the late Middle English (ME) and Early Modern English (EModE) period, had reestablished itself as a language that was also

suited to more elevated modes of discourse. Writers expressed explicit views on style, leading to an emerging consensus over the EModE period about the conventions of various genres, and ideas about appropriate registers for certain discourse domains. Lenker also shows how these developments were reflected in syntactic change, with adverbial connectors and logical linkers shifting from clause-initial to clause-medial position (Lenker 2010: 233ff).

One of the hallmarks of oral versus written styles is the way clauses are connected. The development of a written style tends to involve a tighter syntactic organization: instead of the loosely organized string of main clauses ('parataxis') characteristic of oral styles, written styles tend to have complex sentences, with embedding ('hypotaxis') of subclauses that function as subjects, objects or adverbials of a higher clause.[1]

Example (1) shows a left-dislocation in present-day English (PDE), a sentence beginning with an NP (*The people who earn millions and pay next to no tax*) that is connected to the following clause (*those are our targets*) by the demonstrative *those*, which refers back to the NP. This configuration is paratactically rather than hypotactically organized: the NP has no syntactic function in the actual clause.

(1) The people who earn millions and pay next to no tax, those are our targets. (Birner & Ward 2002: 1413)

Such paratactic constructions are very frequent in Old English (OE). An example is (2), where the clauses and phrases are connected by time adverbs (*Siððan* 'afterwards, then', *þa* 'then'), in bold; note that the punctuation, which influences our interpretation of what is a subclause and what a main clause, may not reflect that of the manuscript and is very likely to have been added or interpreted by the editor:

(2) **Siððan** wæs se III dæg faraones gebyrtyd;
Then was the third day Farao's birthday;
þa worhte he mycelne gebyrscipe his cnihtum;
then prepared he great feastACC his servantsDAT
þa amang þam **þa** geþohte he þara
then among those then remembered he theGEN
byrla ealdor; & ðæra bæcestra.
cup-bearersGEN head and theGEN bakersGEN
'Then on the third day it was Farao's birthday; he then had a great feast prepared for his servants; it was then, among his servants, that he remembered the head of the cup-bearers and the head of the bakers' <Gen (Ker) 44.10>[2]

Comparing the various stages of OE, ME and EModE, the relative numbers of such paratactic constructions can be seen to go down (Figure 11.1).

[1] For the general problem of defining the subclause/main clause distinction on the basis of morpho-syntactic criteria that are cross-linguistically valid, see Cristofaro (2003).

[2] The reference to an OE text enclosed in <> follows the system of short titles as employed in Healey and Venezky (1985 [1980]), in turn based on the system of Mitchell, Ball and Cameron (1975, 1979).

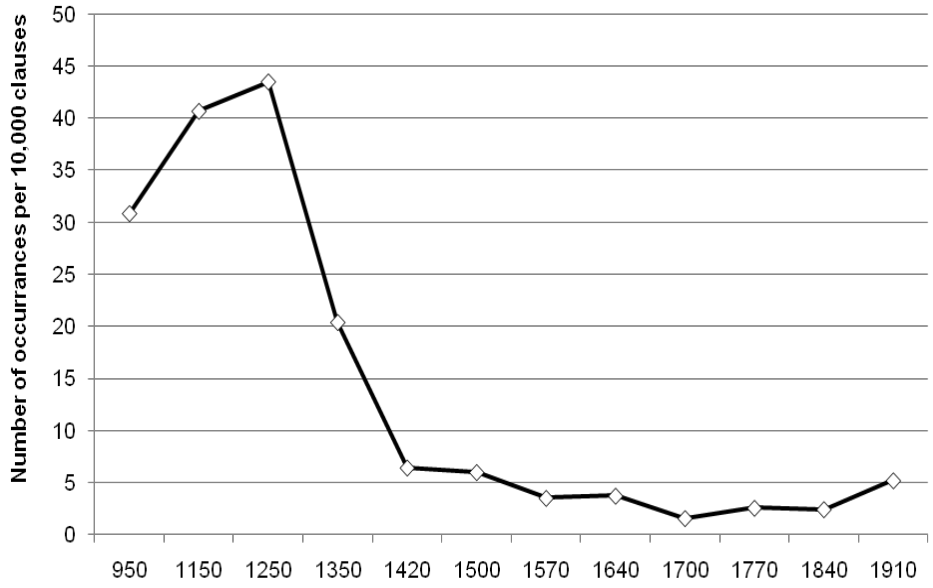

Figure 1. Demonstrative elements in dislocates (Los & Komen in press)

The question is whether such figures reflect genuine language change or whether they are the result of the development of written conventions. They probably reflect both. Written conventions tend to have tighter restrictions on what elements can be elided, and on coreference relations between elements, and tend to rely to a larger extent on explicit syntactic constructions like clefts to meet information-structural needs that may be met by prosodic means in oral styles (see Pérez-Guerra in press). A comparison of oral and written PDE shows that left-dislocations are a feature of spoken rather than written styles (Gregory & Michaelis 2001), and the fact that such written conventions developed in the course of ME and EModE underlies some of the decline in the graph. However, as the construction in (1) is now only possible with subjects, marginally with objects, but not with adverbials such as *then* as in (2) or with adverbial clauses, , we can assume that the decrease in numbers shown in Figure 11.1 also reflects language change.

It is not only emerging conventions for written styles that may obscure investigations into language change, but also the rise of pro- and prescriptivism as a consequence of higher levels of literacy and education. The 18th century was a time of increased social mobility in England: with education becoming more widely available, there was a growing need for normative grammars to help speakers acquire the socially prestigious variety. The existence of socially prestigious varieties has interesting consequences for language change, because speakers who are trying to acquire a language or lect after childhood find it harder to acquire its more subtle phonological, morphological or syntactic aspects, and may hypercorrect, overshoot their mark; An example is the hypercorrect phrase *between you and I* which is the side-effect of speakers trying to avoid the non-prestigious *and me* as the second conjoin of a subject (*Peter and me went to the cinema*), or hypercorrect *whom* in *Whom shall I say is calling*? (Lasnik & Sobin 2000).

For the study of language change on the basis of historical texts, this means that investigators should be aware of the existence of *pro-* and *prescriptive norms*. If there are periods in which the relativizer *which* is felt to be more correct in formal written styles than *that*, or in which pied piping (*the hotel in which I stayed*) is felt to be more correct than preposition stranding (*the hotel which I stayed in*), it is quite likely that a comparison of the various subperiods in a diachronic

corpus shows frequent preposition stranding in ME, more pied piping in EModE, and, perhaps, an increase in stranding in PDE. Such fluctuations should be interpreted correctly, i.e. not necessarily as linguistic change, but as the waxing and waning of the influence of a prescriptive rule. The rule against stranding is well-known, as is the ban on split infinitives, double negation, *who* for *whom*, etc., many of which can be identified in meta-linguistic commentaries of the time. However, there are also rules that do not have such high profiles, and are not taught in school, so that general public and linguists alike are not aware of its existence. Some Dutch publishing houses stipulate, for instance, that in subclauses, the so-called "red" order of past participle and perfect auxiliary (the Dutch parallel of English *has done*) is the only one allowed, not the "green" order (*done has*) because the latter is supposed to be a German word order. The results of any investigation based on texts that are affected by more subtle prescriptive pressures may easily lead to conclusions about language change that are not in fact correct. This is why diachronic investigations always need to pay attention to any variation or exceptional cases in the data that might turn out to be correlated with certain registers or genres - they might be a sign of pro- or prescription rather than a reliable guide to actual language change. We will come back to this in Section 3.

This emphasis on the differences between oral and written styles does not mean that data that show an apparent linguistic change rather than a genuine change are worthless: investigations into change in the spoken or the written language are equally of interest to various research agendas. The point is that data need to be interpreted correctly, and should be embedded in an informed scenario of change.

3. The use of texts for historical sociolinguistics

The previous section addressed the problem of inferring potential differences between written and spoken language on the basis of (historical) texts. Such problems are further compounded when analysing the nature of language variation on the basis of historical texts, with the question of dialectal or sociolinguistic variation in mind. This type of research question requires a social analysis of any particular text: From what dialect area is the author of the text, or the characters featured in the text? How accurate are literary representations of dialect likely to be? What is the author's social status and education? What is the purpose of the text and what is its intended audience/readership? What is the genre and the register of the text within its particular historical context?[3]

The researcher needs to be able to compare texts from different regions, genres and social backgrounds to arrive at at a comprehensive picture of the nature of the variation in the corpus. For the problem of recognizing *dialect variation* in texts we can appeal to a long-standing philological tradition of textual study, in which an inventory of the dialectal characteristics of particular texts has been a central concern. That it is possible to analyse sociolinguistic variation on the basis of texts was suggested early on by Romaine (1982), and there is now a considerable literature on historical sociolinguistics. (See, for example, Nevalainen and Raumolin-Brunberg (2003) for English; Ayres-Bennett (2004) for French; Vandenbussche (2006) for Flemish.) These works provide extensive methodological pointers for the sociolinguistic study of historical texts.

[3] The terms 'genre' and 'register' require some clarification here. In general, the term 'genre' refers to text types such as fiction, essays, letters, drama. The term 'register' is generally reserved for the degree of formality of texts. The two are closely interwoven, of course.

One of the most important of these is to come to a *representative corpus* of text samples, with appropriate *metadata* representing the socially relevant characteristics of the texts. Nevalainen and Raumolin-Brunberg (2003) present many quantitative case studies on the basis of their *Corpus of Early English Correspondences*, tracking the social factors underlying language change in 16th and 17th century English. Ayres-Bennett (2004) focuses on a qualitative analysis of social distinctions in language use in seventeenth century France, basing her observations on the relation between the representation of language in texts and a corpus of metalinguistic commentaries, including observations on the French language, dictionaries, grammars and commentaries, didactic works, treatises on pronunciation, orthography and versification. Vandenbussche has compiled many studies on the characteristics of texts produced by lower-class writers in 19th century Flanders, e.g. Vandenbussche (2007).

Corpus-based research into *register variation* has provided ways in which written and spoken texts can be compared in linguistic terms across time (see e.g. Biber & Finegan 1992, 1994). Biber and Finegan identify three dimensions of textual characteristics, correlating with the oral/literate character of texts: informational vs. involved production, elaborated vs. situation-dependent reference, and abstract vs. non-asbtract style, where in each case there is a scale from formal written (e.g. essays) to informal spoken text (e.g. dialogue in drama). These correlate with a range of linguistic features, allowing an index for each text.The reader may note that these can be matched with a social analysis of the texts involved. On this basis, it has been established in historical sociolinguistics research that, for instance, personal letters pattern more like conversation and drama than other written genres such as fiction, essays, and medical and legal prose. Thus, comparing register/genre differences with what is known about such differences in synchronic states (i.e. in PDE) can be highly instructive, as they set a potential benchmark for differences between genres in historical texts (which of course cannot be taken as absolute; for instance, in English, personal letters as a genre have been treated as close to literary or public writing by some writers at some points in history, and as close to casual conversation by others at other times).These methods are also helpful for an analysis of texts for which the social analysis is absent. Warner (2006) is an example of this: to mine an existing database that was not structured according to genre variation, he uses the criteria of average word length and type-token ratio in order to arrive at a characterization of the stylistic level of 16th and 17th century texts that plausibly corresponds with the oral-literate contrast.

These new approaches to extrapolating historical social contexts can feed into other related areas of study. Establishing sociolinguistic subtleties in a historical corpus can feed back into more formal analysis by helping identify those genres and registers most likely to resemble spoken language. The analysis of text genre and register also draws on traditions in literary studies, and in turn provides a linguistically grounded methodological basis for them.

4. Electronic text corpora

It is probably not an exaggeration to say that the study of textual data was revolutionized in the computer age, as a result of massive *digitization* of written texts: even basic concordance software now allows the researcher to comprehensively search for particular lexical strings, including spelling variants, dialect features, lexical collocations, and with a direct link to the context in which they occur, in a size and type of corpus of the researcher's own choosing, as long as it is digitized text. This alone allows us to study data in a less error prone way and with a comprehensiveness that was hitherto unimaginable. The previous section makes it clear that, in

order to do quantitative sociolinguistic work in particular, a systematic corpus of texts is indispensable.

Many digitized corpora have been morphologically and syntactically *annotated*, allowing more focused explorations into the language of earlier historical stages, in tandem with advances in linguistic theory. Corpora that are syntactically parsed and cover various historical stages exist for English, Faroese, French, Icelandic and Portuguese, with several more in the making for Dutch, German, and Chinese, among others. The availability of such large databases has inevitably changed working practices: the historical linguist no longer has to trawl laboriously through editions, making human errors along the way; the data that took months or even years to collect can now be called up in an afternoon, by the judicious use of search software and the formulation of search queries. The new method poses its own challenges: are the queries correct? Do they find what they are supposed to find? If the query refers to morphological and/or syntactic tags, has the researchers made sure that those tags cover exactly what they intend to investigate? Researchers cannot rely on the bare numbers thrown up by the queries, but need to check the search files in order to make sure that no data are included that should be excluded, or excluded that should be included (see Chapter 13 for more detail on the use of corpora). Researchers need to be aware that electronic corpora are only as good as the texts on which they are based. Text corpora tend to be based on editions, but do not typically offer the benefit of the editor's footnotes or introduction, which may provide the facts of dating (important if an early manuscript is only available in late copies) and indicate where the text stops and the editor's interpretation takes over. Case-endings may have been tacitly expanded from flourishes and diacritics in the manuscripts; passages may have been expanded by fragments from other poems where the editor has added a beginning to the poem from some other source. Earlier (19th and early 20th century) editorial practices went so far as to deliberately "archaise" the text. Punctuation is usually inserted in editions to help the reader, but they may also wrongfoot the reader; see Mitchell (1980) for examples. (For a history of punctuation, see Parkes 1992.)

Researchers need to be aware of such limitations of electronic corpora. However, electronic corpora simultaneously offer many advances in text analysis. One such advantage is the ease of identifying statistical outliers. In Los' (1999) investigation into which verbs could take a *to*-infinitival complement in OE, the example in (5), one such statistical outlier, occurred: it was the only example in the data collection of Callaway (1913) of the verb *cunnian* 'try' being followed by a bare rather than a *to*-infinitival complement (*cunnian* and its bare-infinitival complement in bold):

(5) uton **cunnian**, gif we magon, þone reþan wiðersacan on his geancyrre
let-us try if we may the cruel enemyACC on his return
gegladian <ÆCHom I 30 450.18>
appeaseINF
'let us try, if we can, to appease the cruel enemy on his return'

Callaway followed the punctuation of the edition (Thorpe 1844-1846), which, judging by the commas in (5), took *gif we magon* as a complete clause with "comment" status, an embedded interruption distinct from the syntax of the main clause. This interpretation would imply that the bare infinitive *gegladian* is the complement of *cunnian*. However, the original manuscript has no punctuation at all in this sentence (Clemoes 1955-1956, cited in Healey & Venezky 1985). The availability of the electronic Toronto corpus, not tagged or parsed but containing almost all the

surviving OE texts) made it easy to call up all instances of the verb *cunnian* to see whether they would shed light on the interpretation of (5). About 75% of all occurrences of *cunnian* in the Toronto Corpus are followed by an indirect question with *gif* 'if' or *hwæþer* 'whether'. In the absence of any other attestations with a bare infinitival complement, (5) is in fact best interpreted as yet another such indirect question, with the reading 'let us try/test whether we can appease the cruel enemy on his return', where *gegladian* is not in fact a complement of *cunnian*. The unexplained outlier in the *to*-infinitival data is thus accounted for with the help of new electronic corpus data.

An important side-effect of the use of corpora is that standards set by peer review have become more demanding. As data-gathering can now be done quickly, thanks to corpora and search software, the value of a paper is determined by the quality of the analysis and interpretation of the data rather than by presentation of the data alone. Peer reviewers usually have access to the same corpora, and are able to check the results claimed in a paper, again resulting in higher standards.

5. Caveats and pitfalls

One of the most important messages in studying language in texts, especially over time, is that we must establish standards of *comparability*. This caveat holds for genre and register, as discussed in section 2, but it also applies to comparing texts of the same dialectal provenance, or to distinguishing between competence and performance data. We discuss some examples of this below.

5.1 Comparing like with like: dialect, register, genre

With respect to register and genre, the text material available for various historical stages is often quite diverse (including e.g. for English various kinds of poetry, legal documents, homilies, saints' lives, prescriptive grammars, inscriptions, translations from Latin) and it is difficult to find texts suitable for comparison across historical periods. For instance, OE texts are mostly formal, written in the OE literary language, and are influenced to varying degrees by Latin, directly in the case of glosses and translations, or indirectly as in homilies and saints' lives. In the case of poetry, they may also be influenced by the ancient habits and constraints of the Old Germanic alliterative four-stress line. The language adopted in these genres is different and sometimes hard to compare with that of the Middle English (ME) texts, which comprise, for instance, a rich array of colloquial poetry, and other religious texts beside homilies. The dialect in which most OE texts are written, the West-Saxon *Schriftsprache* (Southern), is only sparsely represented in the extant texts of the Early ME corpus, because few texts from the south in that period have survived. Most ME texts are from the midlands or the north. There is therefore no dialect continuity, and any change we see in a comparison between Old and Middle English texts (for instance, any comparison of the last OE subperiod and the first ME subperiod) may not have been as drastic or as quick as the data suggest. The syntax of the southern dialects appears to be more conservative than that of the Midlands or the North, which means that the rate of loss of Object-Verb order, or the Particle-Verb order, tend to be assessed as fairly steep, suggesting that the change was quicker than was in fact the case. This problem is practically universal in historical linguistics: the balance of wealth and power in the Middle Ages tended to shift from region to region, and so most texts were produced in region A in one period, and in region B in the next. Furthermore, the

survival of manuscripts is subject to the vagaries of history, rendering a degree of arbitrariness. The best a researcher can do is be explicit about data gaps or genre mismatches in their work.

5.2 Comparing like with like: competence and performance

Diachronic investigations have to work with what is known as *performance data*, actual written language use rather than *competence*, the native speaker's internalized grammatical system that allows him or her to construct sentences and judge them on their acceptability (see Chapter 3). The relation between performance (whether written or spoken) and competence has been the object of systematic study to some extent only in sociolinguistics, so we have to be very careful in drawing conclusions about the extent to which the historical texts reflect the grammars of the native speakers who produced them. The most obvious issue here is the question of *negative evidence*. If a construction is not attested in texts of an earlier period, does this mean it was structurally impossible? Again, the situation boils down to comparing like with like: if the relevant structures cannot be found in a synchronic (PDE) "performance" corpus either, even if PDE speakers have no problem constructing them by introspection, the chances are that we are not comparing like with like, i.e. we are comparing performance data from earlier periods to present day competence data. We present some case studies as examples.

The OE text corpus is sufficiently large to allow at times categorical statements of the type *only NPs with accusative case can passivize* (cf. Russom 1983) or '*to' is part of the infinitival phrase and cannot be moved* (cf. Fischer 1996), especially if these phenomena are further confirmed by crosslinguistic evidence from related, living languages. The subsequent rise of passivization of dative NPs, or the splitting of *to*-infinitives, represent ME innovations and have come to be considered as evidence of language change. But unattested structures cannot always be taken as evidence of absence or of diachronic change. Mittwoch (1990:107-108) discusses the difficulties of assessing the status of negation in accusative-and-infinitive constructions, e.g. in examples from introspection such as the sentence in (6) (=Mittwoch's example (33), slightly adapted):

(6) John saw Mary/her not leave

Constructions such as (6) combine an object NP (*Mary/her*) and a bare infinitive (*leave*), and occur after verbs of perception (like *see*) and certain verbs of causation (like *let* or *make* in PDE). Mittwoch makes the point that negated accusative-and-infinitive constructions in PDE are at best "borderline, denizens of some limbo region between the grammatical and the deviant" and adds that, in five years of looking out for real-life utterances, she never encountered a single example, "not even one meant ironically"(Mittwoch 1990:108). This illustrates the gap between performance data and those constructed by introspection. Both have their own valuable contribution to make: the corpus will yield information about usage that might not surface in the laboratory, whereas the laboratory will yield information about structure that might not surface in a corpus study (whether they complement each other completely is a different matter; the extra information produced by each probably does not fully compensate for the other's blind spot). A similar point could be made about the Accusative-and-Infinitive construction with *to*-infinitives after verbs of thinking and declaring, where scholars construct grammatical examples like *I believe them to have a dog* (eg. Miller 2002:149), but also need to account in some way for the

fact that such sentences tend not to show up in performance corpora where the construction occurs overwhelmingly in the passive, and is restricted to quite formal registers (eg. Mair 1990).

The nature of the surviving text material makes it often difficult to find data of the subtlety required for many kinds of analyses. For example, van Kemenade (1987) and Koopman (1990) show that we can get some interesting insights into OE word order if we analyse sequences of verbs in embedded clauses as verb clusters, essentially morphological units. Two examples are given in (7):

(7) a. þæt hie gemong him mid sibbe *sittan* *mosten* <Or 8.52.33>
that they among themselves in peace settle must
'that they must settle in peace among themselves'

b. ðæt he Saul ne *dorste* *ofslean* <CP 28.199.2>
that he Saul not dared murder
'that he didn't dare murder Saul'

This analysis is modelled on analyses for similar verbal clusters in modern German and Dutch, as exemplified in (8a) and (8b) respectively:

(8) a. dass der Johann das Büchlein *haben* *wollte*
that John the booklet have wanted

b. dat Jan het boekje *wilde* *hebben*
that John the booklet wanted have
'that John wanted to have the booklet'

If such an analysis in terms of verb clusters is appropriate, we expect to find further parallellisms. For instance, German and Dutch have long verbal clusters as in (9a) and (9b) respectively.

(9) a. weil er die Kinder *singen* *hören* *können* *hat*
because he the children sing hear can has

b. omdat hij de kinderen *heeft* *kunnen* *horen* *zingen*
because he the children has can hear sing
'because he could have heard the children sing'

Such long verb clusters do not appear in the OE texts. Their absence might reflect their ungrammaticality in OE, in parallel with German and Dutch. However, the absence may be due to rarity in the naturalistic use of this construction. Once again, the availability of corpora now makes it possible to check how frequent these clusters are in the written present-day languages. Coupé and van Kemenade (2009) show that they are generally absent in the full Old West Germanic and Gothic textual record even though they develop in the Dutch language area from the 13th century onward, which would seem to indicate on comparative grounds that they do not form clusters in OE in the way that they do in present-day German or Dutch. But the simple fact is that we have no direct evidence as to the grammatical status of of verb clusters in OE.

These and other cases show that we must always be aware of the strengths as well as the limitations of a corpus of performance data.

5.3 Using data from the secondary literature

When investigating any set of facts, it is useful and necessary to turn to handbooks and other existing literature first. There is a massive amount of literature based on a substantial body of text research, even predating the corpus revolution. One example of this is Visser's (1963-1973) monumental *An Historical Syntax of the English Language*, which includes much of his database. This database needs to be mined with caution (see also Denison 1993: 5). For instance, Lieber (1979) and Lightfoot (1980) claim that OE has indirect passives on the basis of Visser's faulty examples (which crucially leave out dative case markers on the relevant NP, as pointed out by Russom 1982 and Mitchell 1979). Visser's strength lies particularly in the periods after OE; his OE examples are best checked separately, as they include evidence from interlinear glosses and are completely unreliable as a guide to syntactic practice.

There are many excellent late 19th and early 20th Century studies about various syntactic phenomena which include the primary database. A problem that may arise here is that the database may have been originally set up to answer a particular research question, with unfortunate consequences if they are later used to answer different questions altogether. One database that has been extensively mined throughout the 20th Century is Callaway's *The Infinitive in Anglo-Saxon* (1913). Brinton (1988) consults it to find out whether the OE verb *onginnan* 'begin' is showing signs of grammaticalization, in view of the fact that its Middle English reflex *gan* has grammaticalized into an auxiliary, its meaning bleached from 'begin' to something akin to the meaning of the PDE auxiliary *do*.

(19) Witodlice...ongann se hiredes ealdor to agyldenne þone pening
truly began the householdGEN elder to pay the penny
<ÆCHom II, 5 46,137>
'Certainly repaid (*began to repay) the elder of the house the penny' (Brinton 1988:160)

She concludes that *onginnan* cannot mean 'begin' in this OE example, either, because the situation is punctual. The sentence in its entirety is, however, (20):

(20) Witodlice fram ðam endenextan ongann se hiredes ealdor
truly from the last-ones began the houseGEN elder
to agyldenne þone pening. <ÆCHom II, 5 46,137>
to pay the penny
'Truly, from the last ones began the lord of the household to pay the penny.'

The problem is that Callaway, for reasons of space, omitted an indirect object, *fram ðam endenextan* 'from the last ones,' whose plurality would crucially have demonstrated that the event described by the infinitive is iterative and therefore durative rather than punctual.

6. Making the best of data gaps

Linguists working with texts, for instance for the study of language change, genre comparison, or dialect comparison, have to make do with those texts that have survived the vicissitudes of time. The record may not always yield what we want: texts from crucial areas and from crucial periods may be missing from it. The texts we do have lack several dimensions of the spoken word, and,

of course, any direct access to native speaker competence. We end this chapter with two examples of creative solutions to these problems.

OE has a rule of verb placement similar to that in Modern Dutch and German, but with an important difference: with specific types of first constituent, the finite verb (in bold in (13)) will always immediately follow in second position, as in Modern Dutch or German, whether the subject, in third position, is nominal or or pronominal (as *he* in (13)); see van Kemenade (1987):

(13) Þa **gemette** he ðær ænne þearfan nacodne <ÆLS (Martin) 61– 62>
then met he there a beggar naked
'Then he met a poor man, naked'

However, with other types of first constituent, like *Æfter þysum wordum* 'after these words' in (14), subject nominals are still in third position, but pronouns are not: they precede the finite verb, which now looks to be in third place (in bold):

(14) Æfter þysum wordum he **gewende** to þam ærendracan <ÆLS (Edmund) 83>
After these words he turned to the messenger
'After these words he turned to the messenger'

Kroch, Taylor & Ringe (2000) make a case that Northern Middle English, due to language contact with the Scandinavian invaders in the late OE period, only had constructions of the type in (13). Kroch, Taylor and Ringe use 10th century Northern glosses (i.e. interlinear translations, which are generally assumed to be unreliable as evidence) as indirect evidence: where the Latin original does not spell out pronominal subjects, the OE gloss must add them, and this is done in the word order as in (13) rather than (14). They argue on the basis of this fact that in the North, the contact situation with Old Norse (which like Dutch and German has V2 as in (13)) may have affected the verb-second rule directly. This creative use of an atypical data source helps address a particular problem arising out of gaps in the OE record.

The problem of not having access to spoken data is circumvented in Getty (2000). The grammaticalization of (pre)modals, from lexical verbs into auxiliaries, can be expected to have been accompanied by the usual grammaticalization phenomena: bleaching of semantic content, loss of stress, phonetic reduction. Poetry, as a rule, is not used in syntactic investigations for a number of reasons: archaic structures tend to persist in poetry beyond their shelf life in the spoken language, and the requirements of rhyme and metre may also skew the results. However, Getty argues on the basis of the metrical nature of OE poetry that premodals grammaticalize to some extent between early and late OE: they are significantly less likely to occur in stressed positions in the Late OE *Battle of Maldon* than in other, undatable but presumably older, poetry.

7. Conclusion

We have seen in this chapter that working with texts, in particular historical texts, raises a number of specialised issues that require specialised treatment. These may be summed up generally in one question: how and to what extent does the text (or collection of texts) yield the answers to the research question, or, perhaps, how can we make it yield the best possible answer to the research question? We have addressed a range of issues that bear on this question, boiling

down to the representativeness of the textual evidence for the type of information we may wish to draw from the texts.

References:

Ayres-Bennett, W. (2004). *Sociolinguistic Variation in Seventeenth-Century France: Methodology and Case Studies*. Cambridge: Cambridge University Press.

Ball, C. N. (1991). *The Historical Development of the It-Cleft*. Ann Arbor: ProQuest/UMI Dissertation Services.

Bauman, Richard (1986). *Story, Performance, Event: Contextual Studies of Oral Narrative*. Cambridge: Cambridge University Press.

Biber, D. & E.Finegan (1992). The linguistic evolution of five written and speech-based English genres. In: History of Englishes: New Methods and Interpretations in Historical Linguistics, edited by M. Rissanen, O. Ihalainen, T. Nevalainen, & I. Taavitsainen, 688-704. Berlin: Mouton de Gruyter.

Biber, D. & E. Finegan (1994). Introduction: Situating Register in Sociolinguistics. *Sociolinguistic Perspectives on Register*, edited by D. Biber & E. Finegan, 3-12. New York and Oxford: OUP.

Birner, B. & G. Ward (2002). Information Packaging: Chapter 16 in *The Cambridge Grammar of the English Language*, ed. by R. Huddleston & G.K. Pullum, 1363-1427. Cambridge: Cambridge University Press.

Brinton, L.J. (1988). *The Development of English Aspectual Systems: Aspectualizers and Post-verbal Particles*. Cambridge: Cambridge University Press.

Callaway, M. (1913). *The Infinitive in Anglo-Saxon*. Washington D.C.: Carnegie Institution of Washington.

Coupé, G. & A. van Kemenade (2009). Grammaticalization of modals in English and Dutch: uncontingent change. In: *Historical Syntax and Linguistic Theory*, edited by P. Crisma and G. Longobardi. Oxford: Oxford University Press, 250-270.

Cristofaro, S. (2003). *Subordination*. Oxford Studies in Typology and Linguistic Theory. Oxford: OUP.

Delbrück, B. (1882). *Introduction to the Study of Language: A Critical Survey of the History and Methods of Comparative Philology of Indo-European Languages*. Tr. of Einleitung in das Sprachstudium (Bibliothek indogermanischer Grammatiken ; Bd. 4). Breitkopf und Härtel: Leipzig.

Denison, D. (1993). *English historical syntax*. London: Longman.

Fischer, O.C.M. (1996). The status of *to* in Old English *to*-infinitives: A reply to Kageyama. *Lingua* **99**, 107-133.

Fleischman, Susan (1990). *Tense and Narrativity: From Medieval Performance to Modern Fiction*. London: Routledge.

Getty, M. (2000). 'Differences in the metrical behavior of Old English finite verbs: Evidence for grammaticalization.' *English Language and Linguistics* 4, 37-67.

Goody, Jack & Ian Watt (1968). The consequences of literacy. In: *Literacy in Traditional Societies*, edited by Jack Goody, 27–68. Cambridge: Cambridge University Press.

Gregory, Michelle L. & Laura A. Michaelis. (2001). Topicalization and Left Dislocation: A Functional Opposition Revisited. *Journal of Pragmatics* 33: 1665-1706.

Healey, A. D. & R. L. Venezky (1980[1985]). *A microfiche concordance to Old English*. Toronto: The Pontifical Institute of Mediaeval Studies.

Kemenade, A. van (1987). *Syntactic Case and Morphological Case in the History of English*. Dordrecht: Foris.

Koopman, W. (1990). *Word Order in Old English*. Dissertation, University of Amsterdam.

Kroch, A., A. Taylor & D. Ringe (2000). The Middle English verb-second constraint: A case study in Language Contact and language Change. In: *Textual Parameters in Older Languages*, edited by Susan C. Herring, Pieter van Reenen & Lene Schøsler, 353-391. Amsterdam/ Philadelphia: Benjamins.

Labov, W. (1994). *Principles of Linguistic Change*. Vol 1: *Internal Factors*. Oxford: Blackwell.

Lasnik, H. & N. Sobin (2000) 'The who/whom-puzzle: On the preservation of an archaic feature.' *Natural Language & Linguistic Theory* 18, 343-371.

Lenker, Ursula (2010). *Argument and Rhetoric: Adverbial Connectors in the History of English* (Topics in English Linguistics 64). Berlin/New York: Mouton de Gruyter.

Lieber, R. (1979). The English Passive: An Argument for Historical Rule Stability. *Linguistic Inquiry* 10, 667-688.

Lightfoot, D. (1980). The History of NP Movement. In T. Hoekstra, H. van der Hulst, and M. Moortgat, eds., *Lexical Grammar*, Foris, Dordrecht.

Los, B. (1999). *Infinitival Complementation in Old and Middle English* (LOT Dissertation Series 31). The Hague: Thesus.

Los, B. (2005). *The Rise of the* To-*Infinitive*. Oxford: Oxford University Press.

Los, B. & E. Komen (in press). Clefts as resolution strategies after the loss of a multifunctional first position. In: *Rethinking Approaches to the History of English*, edited by T. Nevalainen & E. C. Traugott. New York: Oxford University Press.

Mair, C. (1990). *Infinitival Complement Clauses in English: A Study of Syntax in Discourse*. Cambridge: Cambridge University Press.

Miller, D. G. (2002). Nonfinite Structures in Theory and Change. Oxford: Oxford University Press.

Mitchell, B. (1979) F.Th. Visser, *An historical syntax of the English language:* some caveats concerning Old English. *English Studies* 60, 537-542.

Mitchell, B. (1980). The dangers of disguise: Old English texts in modern punctuation. Review of English Studies n.s. 31, 385-413.

Mitchell, B. (1985). *Old English Syntax*, 2 vols. Clarendon: Oxford.

Mitchell, Bruce, Catherine Ball & Angus Cameron (1975). Short titles of Old English texts. *Anglo-Saxon England* **4:** 207-21.

Mitchell, Bruce, Catherine Ball & Angus Cameron (1979). Addenda and corrigenda. *Anglo-Saxon England* **8:** 331-3.

Mittwoch, A. (1990). On the distribution of bare infinitive complements in English. *Journal of Linguistics* 26, 103-131.

Nevalainen, T. & H. Raumolin-Brunberg (2003). *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. London: Longman.

Olson, David R. (1977). From utterance to text: The bias of language in speech and writing. *Harvard Educational Review* 47, 257–81.

Parkes, M.B. (1992). *Pause and Effect: An Introduction of the History of Punctuation in the West*. Aldershot: Scolar Press.

Pérez-Guerra, Javier. (in press) Discourse status and syntax in the history of English: Some explorations in topicalisation, left-dislocation and *there*-constructions. In Anneli Meurman-Solin, Maria-José López-Couso & Bettelou Los, (Eds.), *Information Structure and Syntactic Change*. Oxford: Oxford University Press.

Perret, Michèle (1988). *Le Signe et la Mention: Adverbes Embrayeurs "Ci," "Ça," "La," "Iluec" En Moyen Français (Xive-Xve Siècles)*. Geneva: Droz.

Plato (1975). *Phaedrus and letters VII and VIII*. Tr. Walter Hamilton. Harmondsworth: Penguin.

Romaine, S. (1982). *Socio-Historical Linguistics: its Staus and Methodology*. Cambridge: Cambridge University Press.

Russom, J.H. (1982). An examination of the evidence for OE indirect passives. *Linguistic Inquiry* 13, 677-80.

Saussure, F. de (1983). *Course in General Linguistics*. Tr. Roy Harris of *Cours de Linguistique Generale*, reconstructed from students' notes after Saussure's death. London: Duckworth.

Traugott, E. C. and M. L. Pratt. 1980. *Linguistics for students of literature.* New York: Harcourt Brace Jovanovich.

Vandenbussche, W. (2007). Lower class language in 19th century Flanders. *Multilingua* 26, 2-3: 279-290.

Visser, F. Th. (1963-73). *An historical syntax of the English language,* Vols. 1-3b. Leiden: E.J. Brill.

Warner, A. (2006). Variation and the interpretation of change in periphrastic DO". In: *The Handbook of the History of English*, edited by A. van Kemenade & B. Los, 45-67. Maldon: Blackwell.