1 English in the digital age: general introduction

IRMA TAAVITSAINEN, MERJA KYTÖ, CLAUDIA CLARIDGE, AND JEREMY SMITH

Modern English historical linguistics is to a large extent an empirical discipline in which language descriptions and theoretical insights are increasingly based on corpus evidence from very large electronic databases. Our aim in this book is to cast new light on how this new approach has transformed our understanding of both contemporary and historical varieties, and to provide an up-to-date, forward-looking account of what is taking place in corpus-based research into the English language. The primary focus of this book is on methodology, and on the use of corpora and other electronic resources to detect language change, or achieve comprehensive synchronic descriptions. But we also hold that these changes need to be related to larger patterns of language history and external sociohistorical developments, and demand moreover an engagement with theoretical underpinnings.

Electronic corpora began to be employed in English linguistics in the 1960s and 1970s, but corpus linguistics really took off on a large scale when computers were reduced in size and became 'personal.' This development took place a couple of decades later, in the 1980s and 1990s, and revolutionized all fields of linguistic research. A key reason for the rapid upsurge of corpus linguistics was, as pointed out by Johansson (2008: 33), that technological advances coincided with a new orientation in linguistics: linguists started to become more and more interested in language use instead of considering abstract language systems (see also below), and computerized materials were quickly found to facilitate access to large bodies of evidence of usage, past and present. First-generation corpora typically included such one-million-word corpora as the Brown corpus (1964, original version) meant to represent language use in the US in the early 1960s; the contents of that corpus were carefully selected to represent distinct genres. The Brown model was followed when the counterpart *Lancaster–Oslo/Bergen* corpus (*LOB*) of British English was created (1976, original version). In English historical linguistics, there had been electronic text collections such as the *Dictionary of Old English Corpus in Electronic Form* (first release in 1981, the current version from 2009), comprising some three million words of Old English and a million words of Latin; the DOEC corpus contains all the

extant Old English texts (represented in the corpus, for the most part, by one manuscript witness). The idea of compiling a large-scale genre-stratified text collection spanning the millennium of the history of English first materialized, however, in the *Helsinki Corpus of English Texts* (1991), and the *Helsinki Corpus* swiftly led to a boom in historical corpus compilation. Subsequent electronic corpora, rather than aiming at representing the whole of the language in its rich genre variation, focused on specialized areas such as letter-writing (e.g. *Corpora of Early English Correspondence*), medical writing (e.g. *Corpus of Early English Medical Writing*) and speech-related texts (e.g. *Corpus of English Dialogues 1560–1760*). And with many corpora compiled today, corpus sizes have increased to hundreds of millions of words (e.g. *Corpus of Historical American English*); such corpora harness the internet to facilitate access to search engines and corpus examples. Indeed, using the World Wide Web as a corpus is one of the recent breakthroughs in the field although – something that affects many aspects of corpus linguistics – there are of course complex issues of copyright and intellectual property which have arisen, not yet resolved.

Today, there are numerous corpus-compilation activities going on: the Brown and LOB corpora have become members of a whole family of corpora, sampling American English from the early 1990s and British English at thirty-year intervals across the twentieth century (BE06, F-LOB, BLOB-1931); new specialized corpora are underway (e.g. newspaper language) and so are huge internet-aided text collections (*Old Bailey Corpus*). Commercial large-scale text collections provide access to vast amounts of printed material covering a wide area of language use such as fiction, drama, poetry, sciences and law across the history of English (e.g. EEBO, ECCO, now being converted to machine-readable form, as EEBO-TCP, ECCO-TCP). Not only printed but also manuscript materials are used in corpus-like ways in electronic text editions (e.g. *An Electronic Text Edition of Depositions 1560–1760*). And of course electronic dictionaries such as the *Oxford English Dictionary* (*OED*) *Online* can in principle be used as corpora, with due caution paid to biases in text representation. Along with the increase in sources, linguists have joined the 'computer-savvy,' eager to learn about new search facilities and new ways of profiting from electronic textual resources. Annotating electronic texts with information on extralinguistic properties such as period, genre and the author's or speaker's social ties, and on intralinguistic properties such as parts of speech, word order, or other language phenomena, increases the value of these resources immensely.

Many of the above-mentioned electronic resources have been made use of in the contributions included in the present volume, and indeed others. Among the resources we have not discussed so far are sources containing spoken British English, e.g. the *Diachronic Corpus of Present-Day Spoken English*, a grammatically analyzed resource which comprises extracts taken from the *London–Lund Corpus* (collected in 1960s–1970s), and the British

component of the *International Corpus of English* (from the early 1990s). Another source of recent British English from the late twentieth century is the 100-million-word *British National Corpus* whose spoken part (10 percent) and written part (90 percent) together comprise texts representative of many different registers and styles. A historical multi-genre corpus comparable to the above-mentioned *Helsinki Corpus* is ARCHER (*A Representative Corpus of Historical English Registers*), which covers the period from 1650 to 1999 and contains specimens of both British and American English. The *Helsinki Corpus* also has a number of satellite corpora which offer files drawn from it in linguistically annotated versions. Of these the *Penn–Helsinki Parsed Corpus of Early Modern English* has been used for a chapter included in the present volume, along with the *Penn Parsed Corpus of Modern British English* which follows as closely as possible its Early Modern English counterpart regarding genre composition. As the compilers of a forthcoming corpus tend to have early access to the materials to be included in their corpus, they are also able to carry out pilot studies to explore the features and potential of the envisaged text collection. The *Corpus of English Religious Prose* is such a project, described and made use of in one of the contributions to the present book. Among the very large-scale text collections exploited for the book are the *Corpus of Late Modern English Texts* (close to 10 million words), comprising e.g. private correspondence, fiction, and scientific prose, the single-genre *TIME Magazine Corpus* (100 million words from the 1920s to the 2000s) and the similarly single-genre *English Drama* component of the Chadwyck-Healey Literature Collections (over 3,900 plays from medieval mystery plays to the early twentieth century drama). Turning to regional varieties, one of the chapters in the book uses material from the one-million-word Singaporean subcorpus of the *International Corpus of English* (conversations, speeches, private letters, academic writing, fiction) and *A Corpus of Singapore Weblogs* (still under compilation). Electronic dictionaries have also been among the sources consulted by the contributors. Of these the online version of *MED* (2001), or the *Middle English Dictionary*, has been around for more than a decade while *EDD Online* is a recent newcomer, based on Joseph Wright's *English Dialect Dictionary* (1898–1905).

The time-span during which such corpora have appeared is short, a span of hardly four decades, but the development – and consequent transformation of the field – has been rapid. Technological advances in search-programs, and the increasing availability of both present-day and historical English language corpora with easy access on the researcher's own computer, have opened up new vistas in English linguistics. We have now come to a new phase where the technical achievements provide a new basis for reassessing the research undertaken using traditional methods, before the advent of electronic resources.

The time is also ripe for us to acknowledge the merits of contextual interpretations, drawing on the sociohistorical background and culture of the period in focus. This approach is in accordance with the latest

developments in English linguistics (see below), but it also pays homage to the discipline's philological back history. As has just been stated, before the advent of corpora, researchers based their conclusions on fewer examples gathered by qualitative reading of older texts. It is of course still possible to find considerable value in (e.g.) the writings of Mustanoja or Jespersen, or indeed of even older scholars such as Sweet, Wyld, or Wright. English historical linguistics, after all, is an accretive subject – like all historiographical endeavors – and new approaches to the subject do not mean that older approaches are simply to be dismissed, even if the idioms in which these theories are expressed are no longer current. For instance, Dr Samuel Johnson, in his preface to his *Dictionary of the English Language* (1755), has some interesting things to say about language change that continue to resonate with scholars. He identifies external causes of change as not just "conquests and migrations . . . now very rare," whose effects are sudden, but also "commerce . . . not always . . . confined to the exchange, the warehouse, or the port, but . . . communicated by degrees to other ranks of the people, and . . . at last incorporated with the current speech" (Bolton 1966: 152). He also identifies "internal causes equally forcible":

> The language most likely to continue without alteration, would be that of a nation raised a little, and but a little above barbarity, secluded from strangers, and totally employed in procuring the conveniences of life . . . men thus busied and unlearned, having only such words as common use requires, would perhaps long continue to express the same notions by the same signs. But no such constancy can be expected in a people polished by arts, and classed by subordination, where one part of the community is sustained and accommodated by the labours of the other . . . Copiousness of speech will give opportunities to capricious choice, by which some words will be preferred, and others degraded; vicissitudes of fashion will enforce the use of new, or extend the signification of known terms. (152–153)

Johnson is here engaging with issues of language contact, social class and stylistic variation; although somewhat unspecific and couched in the distinctive terminology of eighteenth-century linguistics (e.g. "copiousness of speech"), his ideas prefigure quite closely many present-day debates on language variation and change (see e.g. Milroy 1992; Samuels 1972; Smith 1996) – debates which have significant methodological implications.

Dr Johnson also clearly understood the importance of corpora, even if the techniques for developing such materials in the modern manner did not exist for him. His approach to dictionary-making, whereby he established a bank of illustrative quotations as the basis for his definitions, not only prefigured all subsequent lexicographical practice but the development of electronic corpora, e.g. COBUILD, the facility set up by John Sinclair and others from 1980 onwards and which underpinned the *Cobuild Dictionary* (1987). And

of course the *OED*, the greatest of the historical dictionaries which in key ways derived its practice from Johnson's, is similarly founded on the back of a major corpus, i.e. its body of citations.

Traditional language histories, however, are based on qualitative readings of texts, and even in the *Cambridge History of the English Language* (*CHEL*, general editor Richard Hogg), still the most comprehensive treatment extant, very few chapters are based on corpora. It is only the syntax chapters in volumes I to IV which make use of corpora (broadly defined) at all: Traugott (1992) is based on the *Toronto Microfiche Concordance* for Old English and Fischer (1992) uses the *MED* quotations along with a wide selection of texts, but neither of them gives or comments on frequencies in line with current trends in quantitative corpus linguistics. Rissanen (1999) is based on the Early Modern English part of the *Helsinki Corpus*, but makes his frequency statements along the lines of "rare," "common"; only Denison (1998), dealing with Late Modern English, uses corpus(-like) material, e.g. ARCHER, *OED* quotations, Chadwyck-Healey databases, and provides some numerical frequency data.

Denison's approach is logical given the fact that decrease or increase in the use of (variant) expressions signal ongoing change but given the thoroughly attested role of frequency in change (see, for instance, the essays in Bybee and Hopper 2001), the lack of numerical evidence in the other volumes may be seen as surprising. To some extent this state of affairs in *CHEL* can be taken to be an effect of the publication dates. However, in the more recent Blackwell Handbook (van Kemenade and Los 2006) corpus linguistics also plays only a minor role (figuring in only three of twenty-three contributions).

In contrast, the *Oxford Handbook* (Nevalainen and Traugott 2012) has not only a large section on resources (including corpora) but also a section on the use of corpora in observing recent changes. Numerous contributions here as well as in the *Handbook* edited by Bergs and Brinton (2012) make use of corpus evidence; the latter also contains special chapters on corpus-linguistic matters and the importance of frequency studies.

However, in contrast to these newer handbooks which pay tribute to the by now large corpus-linguistic contribution to the field, one-volume textbook treatments of English language history are (still) largely devoid of corpus-linguistic traces. This is partly explicable by original publication date (e.g. Strang 1970), by a more external historical orientation (e.g. Baugh and Cable 2001; Leith 1983; Bailey 1991), by the linguistic orientation of the author (e.g. van Gelderen 2006), or by the fact that the book is intended for absolute beginners (e.g. Culpeper 2005). Brinton and Arnovick (2006), while highlighting the importance of corpora in Chapter 1, do not exhibit the fruits of corpus research in the following chapters, and whereas the exercises make use of the *OED*, electronic corpora (on whatever small scale) are not included. Although "[h]istorical linguistics has been one of the linguistic subdisciplines that benefits most from corpora" (Gries and Hilpert 2012: 134), most

available histories still understate the corpus-linguistic contribution and thus also neglect such fields as historical sociolinguistics, historical pragmatics and genre/text type/register studies (to name but a few), where, for students of modern English linguistics, corpus linguistics has made valuable and innovative contributions.

Harnessing corpora allows a new precision and greater transparency to research practices in the historical study of English; indeed, corpus linguistics allows researchers to place the subject on a firm empirical foundation. In principle, corpus linguistic studies are replicable, so that different researchers should obtain the same result, when using the same methodology and the same corpus. The requirement of replicability, the foundation of the exact sciences, was first formulated in the Royal Society period, by savants such as Thomas Sprat (1635–1713), as the "matter-of-fact" policy (see Shapiro 2003). Thus, for instance, the reliability of data retrieval ensures that fewer data are overlooked that may be potentially relevant for a given research question. Looking for a whole group of spelling variants or a complete semantic field in one search, using wildcards and the like, and clusters with empty or only partially filled slots, allows the researcher to adjust the meshing of the net, from very fine-grained to rather large (see semantic trawling as described by Jucker *et al.* 2012, for example). Both the available search procedures as well as various data-(re)ordering methods offered by software permit the recognition of patterns that might otherwise escape the attention of the researcher. Databases and the like make it easier to link information from various fields, such as intra- with extralinguistic data (e.g. social information).

But new problems arise with the easy availability of large quantities of material in a readily accessible form. Researchers find themselves in a double-edged situation. On the one hand, they demand more rigor in applying statistical methods, and earlier assumptions can now be revised for the overall lines of development or treated in more detail in synchronic descriptions. On the other hand, more contextualization is advocated as changes take place not at some abstract level but in situated language use, i.e. in communication between people. The computer screen can decontextualize examples and shows a very limited view; as a result the data are abstracted from their larger textual contexts. Thus the end-users of corpora may not be familiar with the underlying background facts and sociohistorical developments that are, nevertheless essential for reliable – or at least plausible – results. Moreover, the sources for corpora may be of very mixed and complicated origin; the nature of textual function can constrain, in complex ways, textual form. Thus contexts are needed, especially with reference to early stages of the language where the vagaries of chance survival can easily skew the evidential base.

Context is a complex and multilayered notion that extends from the narrow linguistic context to the broad cultural context. A recent paradigm shift in language studies is the "pragmatic turn," focusing on usage rather than structure, on a linguistics of *parole* rather than a linguistics of *langue* (for a

developed theoretical view see Kretzschmar 2009). Within this new paradigm, pragmatics "has come to be of central interest" and "[v]ariation through space, social group and time, multidisciplinarity, and multiculturalism have all come to play center stage" (Traugott 2008: 207–208). Increasing attention has been drawn to dynamic views of language use, with more flexible, and activity-centered viewpoints gaining ground (see e.g. Blommaert 2005). An even newer trend can be seen in the spread of discursive assessments to phenomena like politeness that were earlier regarded as more stable qualities; in recent studies politeness is constructed in the unfolding discourse with momentary changes in the interaction even in written texts (see Jucker and Taavitsainen 2013).

In the present volume, developments of English are illuminated from different angles and in different time spans, but all make reference to the new insights derived from what seems to be this genuine paradigm shift. Some chapters cover a long diachrony, while others focus on recent decades. A broad range of corpus-linguistic methods is applied on an equally broad range of corpora, and other electronic resources complement the assessments. All chapters of this book make use of electronic evidence, from fully parsed corpora to newly compiled databases and an unusual quasi-corpus of dialect materials. The latter in particular show how the range of electronic sources is expanding, also including more sources with not only a quasi-corpus-linguistic but also a metalinguistic character (see Anderwald 2012 for the *Collection of Nineteenth-Century Grammars*). Prototypical corpora are thus joined by a variety of digital or digitizable sources, which complement each other to produce a more rounded picture of the current situation in English linguistics.

The essays in this volume are designed to capture something of the excitement involved in a major shift not only in the methodology of English historical linguistics but also in its theoretical underpinning. This book has been organized into sections dealing with various aspects of diachronic corpus linguistics and although it will become swiftly apparent that there are numerous cross-themes, we selected the most salient ones as the organizing principles. Part I sets the scene with two chapters on methodology in dialogue with one another about issues of importance to the corpus linguistics approach, and how to apply corpus linguistics for the best possible results. Together these two contributions form a whole, followed by a case study. Part II continues the line of case studies but focuses on changing patterns of syntax and semantics, at the same time illustrating what new corpora and other electronic resources can yield. The first two chapters have several points in common, as discussed in the Introduction to Part I. The second and the third contributions deal with speech-based data, and the latter bridges to Part II, as it deals with interjections. Thus pragmatics is present already in Part II, but Part III is explicitly devoted to the emerging field of corpus pragmatics, showing how 'marginal' language data such as

interjections and hesitators can open up new research agendas. The order here is from micro to macro, as the last chapter of this Part deals with large-scale mapping of religious language use whose influence on the styles of writing in the history of English is acknowledged but somewhat understudied. Part IV is devoted to the diversification of Englishes in different parts of the world where distinct varieties have developed in various contexts and in contact with various cultural influences. It begins with two chapters on language and identity, which is a topic of increasing importance in present-day worldwide Englishes. The third chapter provides a large-scale overview of phonological developments in the varieties, and the speech theme is continued in the last chapter illuminating the latest achievements in the field.

Part I

Linguistic directions and crossroads: mapping the routes