The Routledge Handbook of English Language and Digital Humanities

Svenja Adolphs, Dawn Knight

Written corpora

Publication details
Sheena Gardner, Emma Moreton
Published online on: 05 May 2020

How to cite :- Sheena Gardner, Emma Moreton. 05 May 2020, Written corpora from: The Routledge Handbook of English Language and Digital Humanities Routledge
Accessed on: 13 Dec 2023

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Introduction

The humanities has always been concerned with written texts, but the digital revolution has allowed us to electronically store, annotate and analyse ever-increasing amounts of data in new and creative ways. Indeed, the digital humanities (using digital tools and techniques, including corpus linguistic methods to address humanities-related research questions) offers an opportunity to bring together scholars from across the disciplines to look at ways of harnessing the power of new technologies in the study of humanities data. Here we focus on corpora of written texts.

Collections of written texts, or archives, which are simply assembled and may or may not be indexed and catalogued are differentiated from corpora, which involve a careful sampling of texts that are representative, for instance, of written English and are digitally prepared for analysis. Corpora are generally understood to be ‘collections of naturally-occurring language data, stored in electronic form, designed to be representative of particular types of text and analysed with the aid of computer software tools’ (Nesi 2016: 206). Corpora include not only the actual texts but also metadata, or information about the production of the texts (e.g. who, where, when, why, about what), which is used to associate the text with its social context and to assist in comparative analyses within and across corpora.

Written corpora can be analysed from a range of quantitative perspectives to uncover patterns that would be hard to detect manually. These might be based on frequency, co-occurrence or relative frequency. For example, frequency data might be sought to answer questions such as ‘Which nations do Shakespeare and Dickens refer to most in their writing?’ Or co-occurrence queries might be used for more complex questions, such as ‘Which images are typically associated with darkness in Shakespearean vs. Dickensian writing?’ This latter question might be answered quantitatively in terms of most frequent collocating items, which in turn might simply reflect the nature of English. For example, we can identify which adjectives most frequently occur with a noun such as ‘night’ and compare the co-location findings for Shakespeare and Dickens. Collocation search parameters can be set to count only those that occur immediately before the noun or to include those that occur up to a certain number of words before and/or after the noun. Alternatively, the question could
Written corpora

be answered in relative terms by comparing items that are distinctive, or ‘key’, in the writing of each author when compared with a general English corpus. These central concepts of frequency, collocation and keyness are widely used and explained in all introductions to corpus linguistics (e.g. Hunston 2002). More complex ‘query language’ would be needed to answer questions such as ‘How are women represented in Shakespeare and Dickens?’, and as the results are unlikely to be as directly comparable, an interpretation of the results will normally involve grouping the results into categories, which may be linguistic (e.g. semantic categories or categories of metaphor) or may derive from theories of gender representation or literary criticism.

Research on written corpora involves using tools to access and manipulate the information in corpora in ways that can answer questions, particularly those related to frequency, phraseology and collocation within and across texts, that our intuitions about language or our linear reading of texts cannot. Around this quantitative methodological core, and the more qualitative and interpretive examination of significant or key items (words, phrases or collocations) and the contexts in which they occur, the questions posed and how findings are explained are as varied as research in humanities itself.

Our aim in this chapter is to describe influential written corpora in English. We present a broad categorisation (Table 3.1), followed by an account of historical developments in the field and issues related to corpus contents and the growing statistical sophistication of analytical tools. We then present two small-scale studies: a corpus search for the term *humanities* in several large, ready-made, freely available corpora and an investigation of features of correspondence in three specialist corpora. The chapter concludes with suggestions for future directions, further reading, related topics and references. Our aim is to demonstrate how written corpora, and corpus methods of analysis, complement and contribute to the multidisciplinary field of digital humanities.

**Types of corpus**

The categorisation in Table 3.1 is intended to give a broad overview of types of written corpus that exist in English. Similar classifications are found in most introductions to corpus linguistics (e.g. Huber and Mukherjee 2013; Hunston 2002: 14–16; the Lancaster University corpus linguistics MOOC session 2), as well as on websites that offer access to multiple corpora, such as Brigham Young University (http://corpus.byu.edu) and Sketch Engine (www.Sketch Engine.co.uk/user-guide/user-manual/corpora/corpus-types/).

This classification is not exhaustive; other categories include monolingual (most of the corpora listed are monolingual English), multimodal (e.g. transcripts plus recordings), reference (large general corpora are often used as reference corpora against which smaller bespoke corpora can be compared), pedagogic (e.g. all the receptive texts for a learner – textbooks, lectures, etc.) and regional (the International Corpus of English [ICE] was designed to compare texts from 20 varieties of English, while the SCOTS corpus focuses on a single variety). Nevertheless, it provides an orientation to the scope of written corpora in English.

The earliest national corpora are the Lancaster-Oslo-Bergen (LOB) corpus of British English and the Brown corpus of American English, both developed in projects that began in the 1960s. These were succeeded by the Bank of English in the 1980s (see information on COBUILD later) and other larger national corpora, specifically the British National Corpus (BNC) developed between 1991 and 1994 and the Corpus of Contemporary American (COCA), both of which are widely used today. The BNC is valued for its extensive annotation and multiple categories. It reflects an adult educated variety of English with
Sheena Gardner and Emma Moreton

Table 3.1 Types of corpus exemplified

<table>
<thead>
<tr>
<th>Type</th>
<th>Contents</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General</strong></td>
<td>A variety of texts intended to reflect general uses of spoken and written</td>
<td>The British National Corpus (BNC), a general corpus of British English</td>
</tr>
<tr>
<td></td>
<td>language</td>
<td>The Corpus of Contemporary American English (COCA), a general corpus of</td>
</tr>
<tr>
<td></td>
<td></td>
<td>American English</td>
</tr>
<tr>
<td>**Historical/</td>
<td>From specific periods, or allows comparison by, e.g. decade</td>
<td>Hansard corpus (1803–2005)</td>
</tr>
<tr>
<td>Diachronic</td>
<td></td>
<td>Time Magazine corpus (1923–2006)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The Siena Bologna/Portsmouth Modern Diachronic Corpus (SiBol/Port) contains</td>
</tr>
<tr>
<td></td>
<td></td>
<td>787,000 articles (385 million tokens) from three UK broadsheet newspapers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>in 1993, 2005 and 2010</td>
</tr>
<tr>
<td><strong>Learner</strong></td>
<td>Texts produced by learners</td>
<td>International Corpus of Learner English (ICLE) with 3 million words of</td>
</tr>
<tr>
<td></td>
<td></td>
<td>argumentative essays</td>
</tr>
<tr>
<td><strong>Monitor</strong></td>
<td>Regularly updated to track changes in a language</td>
<td>Cambridge Learner Corpus (CLC) with 50 million words from Cambridge exam</td>
</tr>
<tr>
<td></td>
<td></td>
<td>scripts submitted by 220,000 students from 173 countries</td>
</tr>
<tr>
<td><strong>Parallel</strong></td>
<td>Texts and their aligned translation(s)</td>
<td>The Bank of English (BoE) (see discussion at <a href="http://corpus.byu.edu/coca/">http://corpus.byu.edu/coca/</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td>compare-boe.asp)</td>
</tr>
<tr>
<td><strong>Multilingual</strong></td>
<td>Containing texts in several languages</td>
<td>News on the Web (NOW) corpus dates from 2010 and grows by about 10,000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>web articles each day</td>
</tr>
<tr>
<td><strong>Specialist</strong></td>
<td>Designed for a specific purpose</td>
<td>EUR-LEX is a multilingual parallel corpus of European Union documents</td>
</tr>
<tr>
<td></td>
<td></td>
<td>translated into the official European languages</td>
</tr>
<tr>
<td><strong>Web Corpus</strong></td>
<td>Created by web crawling, categorised by web domain, typically the most</td>
<td>British Academic Written English (BAWE) contains 6.5 million words of</td>
</tr>
<tr>
<td></td>
<td>synchronous of corpora</td>
<td>assessed university student writing across disciplines and levels of study</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hong Kong Professional Corpora specialises in financial services,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>engineering, governance, etc.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>See also ‘Small correspondence corpora’ in the text</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ententen (2013) contains 19 billion words of English on the web</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GloWbE contains 1.9 billion words from websites in 20 different English-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>speaking countries</td>
</tr>
</tbody>
</table>

90 percent written texts, including media, literary and academic sources across domains, and 10 percent transcripts of spoken language. Broadly speaking there are massive corpora compiled by web crawling. For example, Sketch Engine houses TenTen corpora that aim to be more than 10 billion words each for more than 30 languages and a news corpus of 28 billion words that grows by 800 million words a month (www.Sketch Engine.eu/timestamped-english-corpus, accessed 17/04/2018). There are large (COCA had over 560 million words in December 2017) general corpora, such as the BNC and COCA, that collect a range of genres according to a specific design; there are corpora built on existing archives or collections, such as the EUR-LEX, Hansard, Time Magazine and CLC corpora; and corpora that are compiled for specific research purposes, such as the ICLE and British Academic Written
Written corpora

English (BAWE) corpora. Corpora also vary in the nature and extent of their annotation or mark-up, and thus how amenable they are to answering specific questions in the humanities. Our two sample analyses later provide some insights into this, as do other chapters in the handbook.

Background

Research across the humanities has long been informed by the analysis of written texts. Pre-digital studies used techniques that are familiar today. Concordances of texts such as the complete works of Shakespeare and the King James Bible were created manually (see Wisby 1962) by searching for patterns and extracting items with similar functions, references or meanings and published as reference books.

In linguistics, Peter Fries (2010) describes Charles Fries’s systematic collection, scientific analysis and counting of relative frequencies of paradigmatic features (e.g. intonation patterns associated with yes–no questions) in ‘a representative sample of the language of some speech community’. Fries Junior argues that, unlike the transformational linguistics that came to dominate the field particularly in America, the early twentieth-century linguistics of Firth, Quirk, Fries and others was a natural precursor to current corpus linguistics. This explains Thompson and Hunston’s identification of three concerns shared by systemic-functional linguistics and corpus linguistics: a focus on ‘naturally occurring language, the importance of context (or contextual variation), and of frequency’ (2006: 5).

Similarly in folklore and literature studies, Michael Preston (n.d.) describes the development of computerised concordances:

Computer-assisted study of folklore and literature was initiated shortly after World War II by Roberto Busa, S.J., who began preparing a concordance to the works of Thomas Aquinas in 1948, and Bertrand Bronson, who made use of the technology to study the traditional tunes of the Child ballads. In the pre-computer era, Bronson had worked with punched cards which he manipulated with a mechanical sorter and a card-printer. Many early efforts at producing concordances by computer were modeled on the punched card/sorter/reader-printer process, itself an attempt at mechanizing the writing of slips by hand which were then manually sorted.

What have changed are the speed and ease with which we can now collate large quantities of written texts and, when they are digitally transcribed and marked up, the speed and ease with which we can analyse them.

Two of the earliest digital corpus projects were the Brown Corpus of American English (www.helsinki.fi/varieng/CoRD/corpora/BROWN/ accessed 17/04/2018) and its counterpart, the LOB corpus of British English. Developed in the 1960s, these each contained 500 texts of around 2000 words distributed across 15 text categories. This inspired the compilation of further corpora with the same design that enabled ready comparison across varieties of English. Such developments whetted researchers’ appetites to develop bigger and more diverse corpora.

An ambitious project was conducted at the University of Birmingham under the leadership of John Sinclair. The project involved not only the construction of a very large corpus of contemporary English – it is currently around 450 million words (www.titania.bham.ac.uk/docs/svenguide.html) – but also the development of methods and approaches...
to corpus analysis that changed the way we work with English text and still pervade the field today. The annual Sinclair Lectures regularly remind us not only of the achievements of the COBUILD project (see later), and the related advances in lexicography in particular, but also that when it started in the 1980s, it was the English Department, and his project, that had the largest and most expensive computer on campus. The project was funded by Collins and created to inform the making of dictionaries, grammar books and related teaching materials.

In the last 50 years, we have moved from mainframe computers, with data entered on file cards, reminiscent of the handwritten cards used by Charles Fries. In the 1980s these were replaced by desktop computers, which allowed individual researchers to develop and analyse their own digital corpora using concordancing software, and today many hundreds of digital corpora of written texts are increasingly available for online use with sophisticated and user-friendly software programmes by students from mobile devices.

**Critical issues and topics**

Much humanities research involves analysing and interpreting texts, and it is therefore not surprising that researchers are attracted to digital analyses of written corpora to seek evidence in or interpretation of texts to support their theories. There are many decisions to be made in compiling a corpus, however, and when a corpus has been designed for one purpose, it may not be appropriate for use in different contexts. As the number of publicly available corpora increases, the methods used to compile them come under increasing scrutiny. When the COBUILD project began, the aim was to collect large numbers of words, and so novels and academic journals were an easy source of written data. Equally, when an academic corpus was being developed by Averil Coxhead in order to identify items for an academic word list (AWL) (www.victoria.ac.nz/lals/resources/academicwordlist/information/corpus), corpus collection was heavily influenced by the textbooks at one university and is biased more towards business than sciences, for instance. This and other critiques have led to alternative corpora to explore in the production of new AWLs (e.g. see review in Liu 2012). Nevertheless, both the COBUILD and the AWL projects have significantly influenced English language teaching and have inspired the development of more ‘representative’ corpora.

There are, of course, issues of what is ‘representative’, and perhaps a more realistic goal for many is to develop a ‘balanced’ corpus. By identifying particular features of interest in advance (e.g. male vs. female, humanities vs. sciences, quantitative vs. qualitative research, correspondence to subordinates, to superiors, or to peers), it should be possible to balance the corpus in these respects. It will never be possible to have corpora that are balanced for all situational and contextual variables. So, corpus developers should be clear about their research goals when they are developing the corpus, and corpus users should seek to understand not only the composition of the corpus but also the rationale for its construction, which features are ‘balanced’, and in what ways.

A third critical issue relates to the statistical sophistication needed to interrogate and use corpus findings. The statistics exemplified in areas such as multidimensional analysis (Biber 1988) and the use of R (Gries 2013) is crucial to the development of corpus linguistics as a field, but these are not traditional strengths of researchers in humanities, and the extent to which those investigating corpus resources need to understand, or teach, the statistics behind the findings is a moot point.
Current contributions and research

Corpora are used in many areas of English language research. They inform theories of language and of language learning and teaching (e.g. Frankenberg-Garcia et al. 2011), translation, second language acquisition and (critical) discourse analysis. For instance, Caldas-Coulthard and Moon (2010) use corpora to help ‘deconstruct hidden meanings and the asymmetrical ways people are represented in the press’ (2010: 99).

Here we illustrate the scope of written corpora and corpus research through two key domains: diachronic English linguistics and English language teaching materials.

Diachronic/historical corpora

In terms of historical written corpora, a notable starting point is A Representative Corpus of Historical English Registers (ARCHER). Initially constructed in the 1990s by Biber and Finegan (Biber et al. 1994), ARCHER is a ‘multi-genre historical corpus of British and American English covering the period 1600–1999’. Other multi-genre diachronic corpora include the Corpus of Historical American English (COHA), which includes over 400 million words dating from the 1810s–2000s, and the Corpus of English Religious Prose containing a range of genres (including prayers, sermons, treatises and religious biographies), dating from 1150 to the end of the eighteenth century. Genre-specific corpora include the recently released Corpus of Historical English Law Reports 1535–1999 (CHELAR) – 185 files (369 texts) totalling 463,009 words – and the Zurich English Newspaper Corpus (ZEN), consisting of 349 complete newspaper issues containing 1.6 million words.

Arguably, some of the most useful materials for exploring language change and variation are ego documents: diaries, first-person narratives and private correspondence. The Corpus of Early English Correspondence (CEEC), produced at the University of Helsinki, comprises 188 letter collections (12,000 letters) dating from c. 1403 to 1800. The resource serves as an excellent example of how sociobiographic and extra-linguistic metadata might be captured in a systematised way, allowing users to explore language in relation to variables such as the author’s sex, rank, education and religion (Raumolin-Brunberg and Nevalainen 2007: 162). Researchers at VARIENG (the Research Unit for the Study of Variation, Contacts and Change in English, at the University of Helsinki) have used this material to explore, for instance, expressions of politeness, spelling variation and lexical change.

Additionally, a growing number of digital humanities projects are now using letters as a primary data source. With some basic pre-processing work, these digitised correspondence collections can be used for linguistic analysis. For example, the Mapping the Republic of Letters project at Stanford University, in collaboration with various partners including Oxford University’s Cultures of Knowledge project, maps networks of correspondence between ‘scientific academies’ within Europe and America during the seventeenth and eighteenth centuries to explore how such networks facilitated, amongst other things, the development and dissemination of ideas and the spread of political news.

Other projects include the Electronic Enlightenment project, at the University of Oxford, containing over 69,000 letters and documents from the early modern period and the Darwin Correspondence Project at Cambridge University, which has collected and digitised roughly 15,000 letters by Charles Darwin, providing information ‘not only about his own intellectual development and social network, but about Victorian science and society in general’. Finally, the Victorian Lives and Letters Consortium, coordinated by the University...
of South Carolina, has brought to light samples of life-writing from the period spanning the
coronation of Queen Victoria to the outbreak of World War I. Documents include letters by
Thomas Carlyle and the diaries of John Ruskin.¹

Smaller-scale projects include Sairio’s (2009) study of correspondence by Elizabeth
Montagu (as part of her research on letters of the bluestocking network) and Tieken-Boon
van Ostade’s (2010) study of unpublished correspondence of Robert Lowth. Other digi-
tal collections include The Browning Letters – 574 letters between Victorian poets Robert
Browning and Elizabeth Barrett Browning; Bess of Hardwick’s Letters – 234 letters dating
from c. 1550 to 1608 all available in Extended Markup Language (XML) format; and the
Van Gogh Letters Project – around 800 letters by van Gogh to his brother Theo as well as to
artist friends, including Paul Gauguin and Emile Bernard (translated into English).²

Personal and corporate letters of eminent persons (e.g. public political and social fig-
ures) have long been used for social, historical and cultural studies. Such letters are saved,
transcribed, edited and published. Eminent persons themselves are sometimes part of this
process – making copies of their own letters before sending them or retrieving letters from
recipients. The practices of Thomas Jefferson, for example, have received much attention:
his efforts to repair the research record and his use of the letterpress and polygraph letter
duplication systems (see Sifton 1977). Over the past decades, however, there has been a
growing interest in what scholars have termed ‘history from below’ or ‘intrahistoria’³ –
that is, a history of the popular classes – ordinary men and women who experienced his-
torical events and their consequences firsthand (see, for e.g. Amador-Moreno et al. 2016;
Auer and Fairman 2012; Fairman 2012, 2015). For example, the Letters of 1916 project
is creating a crowdsourced digital collection of letters from the time of the Easter Rising
(1915–1931) and the Irish Emigration Database (IED) housed at the Mellon Centre for
Migration Studies in Northern Ireland, which contains over 4,000 letters by Irish migrants
and their families dating from the seventeenth to the twentieth century and has been used
to explore topics and themes in the discourse, as well as linguistic variation and change in
Irish English.

Historical literary corpora are also being used to investigate language change and vari-
ation. The Salamanca Corpus contains literary texts from the 1500s to 1900s searchable
by period, county, genre and author, allowing users to explore vernacular literature and the
representation of dialect; the York-Helsinki Parsed Corpus of Old English Poetry contains
71,490 words of Old English text that has been parsed syntactically and morphologically,
enabling linguistic analyses of various kinds; and the Visualizing English Print project (a
collaboration between the University of Wisconsin-Madison Libraries, the University of
Strathclyde and the Folger Shakespeare Library) has recently released plain-text files of
early modern text and drama corpora. The variant detector VARD, a pre-processing tool
designed to deal with spelling variations in historical data, has been used to standardise the
texts for the corpus analysis work. VARD is also useful in other areas of humanities research
(error tagging in learner corpora (Rayson and Baron 2011), the normalisation of spelling
variations in text messages (Tagg et al. 2012) and the exploration of gender and spelling
differences in Twitter and Short Message Service (SMS) (Baron et al. 2011).

Other literary corpora include the Corpus of English Novels (CEN) – texts by 25 British
and American authors dating from 1881 to 1922, and the Shakespeare Corpus, produced by
Mike Scott, containing 37 plays, including separate files for all speeches by all characters
within those plays. Finally, the Charles Dickens Corpus contains 14 texts (over 3 million
words) and forms the basis of a collaborative project – CLiC Dickens – between the Univer-
sity of Nottingham and the University of Birmingham, which ‘demonstrates through corpus
Written corpora

stylistics how computer-assisted methods can be used to study literary texts and lead to new insights into how readers perceive fictional characters’ (CLiC Dickens).

**English language teaching materials**

A major strand in corpus research involves corporate collaboration to inform English language dictionaries, grammar books and teaching materials. The Collins Birmingham University International Language Database (COBUILD) project (Moon 2007) started by John Sinclair in the 1980s and informed by the Bank of English corpus, produced the COBUILD dictionary in 1987, followed by pattern grammars and teaching materials. Here we see corpus-driven research – corpora being used not to find current examples of grammatical and lexical usage, but to inspire a step-change in lexicography, a novel way of viewing language and a new theory of grammar as patterns. The COBUILD team produced teaching resources that challenged traditional notions of what to teach and how to teach it. For instance, compared to Hornby’s (1974) 25 formal verb patterns, Francis, Hunston and Manning’s pattern grammar (1996) identifies over 700 meaningful verb patterns; Willis (1994) demonstrates alternative explanations of three common grammatical features – the passive, the second conditional and reported statements – and reveals the value of lexical phrases. This interest in lexical phrases (or clusters, or n-grams, or bundles) has become a cornerstone of corpus linguistic research and has inspired new theories about how we learn language (e.g. Hoey’s lexical priming, 2005).

The *Longman Grammar of Spoken and Written English* (Biber et al. 1999) builds on the Quirk et al. (1985) grammars and provides a descriptive grammar of English informed by the 40-million-word Longman Spoken and Written English (LSWE) corpus. One of its distinguishing features is the inclusion of register frequency information in spoken, media, fiction and academic texts. For example, although present-tense verbs are more common than past-tense verbs and are frequent in conversation and academic prose (for different reasons), in fiction past-tense verbs are more frequent (p. 456). Furthermore, some verbs, including *bet, know, mean* and *matter*, occur more than 80 percent in the present tense, while others, including *exclaim, pause, grin* and *sigh*, occur more than 80 percent in the past tense (p. 459).

A subsequent project was supported by the Educational Testing Service (ETS) to investigate the spoken and written registers that university students in the United States listen to or read, including classroom teaching, textbooks and service encounters (Biber et al. 2002). The data in the 2.7-million-word TOEFL 2000 Spoken and Written Academic Language (T2k-SWAL) corpus was collected from four sites across the United States and stratified by level and area of study. This design is well suited to a focus on the language needed to study at university, as assessed in TOEFL.

As these are essentially commercial projects, access to these corpora tends to be restricted, but in some cases it is possible to gain access on request.

Nowadays, many publishers employ their own corpus linguists to work on in-house corpora. Notable examples include Cambridge University Press, Longman, Oxford University Press and Pearson, all of which can use the works they publish in corpora developed for different purposes. For instance, the *Oxford Learner’s Dictionary of Academic English* (Lea 2014) is informed by and includes extracts from the 85-million-word Oxford Corpus of Academic English (OCAE) which includes undergraduate textbooks, scholarly articles and monographs. The freely available Pearson academic collocates list was compiled from the written component of the 25-million-word Pearson International Corpus of Academic English (PICAE), with the support of many well-known corpus linguists.
Other influential academic English corpora include those developed by Ken Hyland and used to develop a theory of metalanguage (Hyland 2005) and to provide evidence that supports arguments in favour of the disciplinary distinctiveness of academic English (and therefore for teaching English for specific academic purposes). Hyland (2008), for example, compares lexical bundles across disciplines in a 3.5-million-word corpus of research articles, master’s and PhD theses from Hong Kong.

Freely available academic corpora include the 6.5-million-word ESRC-funded BAWE corpus of successful university student writing (www.coventry.ac.uk/BAWE). It has not only informed a genre classification (Nesi and Gardner 2012) and studies of academic English, including the use of shell nouns (Nesi and Moreton 2012), reporting verbs (Holmes and Nesi 2009), nominal groups (Staples et al. 2016), section headings (Gardner and Holmes 2009), Chinese writers (Leedham 2015) and lexical bundles (Durant 2017), but it also underpins the Writing for a Purpose materials on the British Council LearnEnglish website (https://learnenglish.britishcouncil.org/en/writing-purpose/).

In this overview of current contributions and research, we have included just some of the written corpora, most of which are accessible online. More corpora are listed in the Further Reading databases and interfaces at the end of this chapter.

Sample analyses

Two contrasting sample analyses of written corpora are now presented.

**Humanities in pre-loaded general, academic and historical corpora**

Here we illustrate the nature of information that can be freely obtained from pre-loaded corpora on the open Sketch Engine and Brigham Young University (BYU) sites through searches for **humanities** in several popular corpora of written texts. The choice of **humanities** is not only appropriate in relation to the contents of this volume but also appropriate in that we consider the term to be a semantic shifter (Fludernik 1991); a term that has forms (signifiers) and referents, but whose meaning (signified) is understood according to the contexts in which it is used. The aim is not to suggest that this is how a study should be conducted, but to show through example the kinds of information that can be obtained from different corpora.

**BNC**

We started with a WordSketch search for the lemma **humanity** on the BNC at www.SketchEngine.co.uk/open. The results show that the only context in which it is frequent is as a postmodifier with ‘of’, where it occurred 432 times (3.85 pmw), compared to the next most frequent context ‘in humanity’ with 87 (0.77 pmw). Examples with ‘of’ include ‘a sense of common humanity’, ‘the dregs of humanity’, ‘the whole of humanity’ and ‘in the interests of humanity’. As in these examples, it is widely (396 times, or 3.52 pmw) used as a singular, uncountable, abstract noun, ‘humanity’, rather than a plural form. Among collocates of the plural form ‘humanities’ is a group of educational terms (**humanities scholars, student, research, department, course**), as in this example, which is worth reading for the ideas it promotes as well as the use of the term ‘humanities’.

What will the next decades hold? Will it be seen, with hindsight, that the introduction of the computer into the humanities disciplines has made no more substantial change in the
eventual output of research than was made by the introduction of the electric typewriter? Or will there be a flowering of research papers, neither pedestrian nor spectacular, containing solid and original results obtained by techniques for which the computer is indispensible? It is hard to make a confident prediction. But the testing time has now arrived; because for the first time posts of leadership in humanities departments are being taken up by a generation of scholars who have been familiar with the computer . . .

Although the use of computers in humanities dates back to the 1960s, this extract from a published lecture shows the state of development in the early 1990s, 30 years later. Through analysis, the BNC can provide an understanding of the meanings, collocations and grammatical patterns of humanity (and humanities) in a large general English corpus from the last century. These two analyses begin to illuminate the term, and much more could be done with further exploration of a wider range of uses and more sophisticated techniques.

**BAWE**

The British Academic Written English (BAWE) corpus allows us to search for humanity by genre family, level of study and discipline to find out how it is used in student writing in the early twenty-first century. If we filter to discipline and search for collocates, we see that it collocates differently in law, philosophy/ethics, English literature and classics, as in these examples.

**‘crimes against humanity’ in law**

individuals at the highest levels of government and military infrastructure could be prosecuted and punished under international law, recognising accountability for the commission of war crimes and introduced new offences of crimes against peace and crimes against humanity.

**‘common humanity’ and ‘humanity can survive’ in ethics/philosophy**

Indeed it is the similarities between so-called cultures that allow us to appreciate that there is a common humanity. If there was not then we would be unable to comment on the cultures of the past or of different location because there would be nothing we could relate to. This is shown in . . .

**‘his/her humanity’ in English literature**

a microcosmic happy ending in Emma prefigures the macrocosmic happy ending of the whole book, for Mr Knightley, being true to his name, moderates Harriet’s embarrassment by dancing with her himself, thus proving his humanity and that he is the ideal husband for Emma.

Nineteen of the 31 instances of ‘humanity’ in the law texts relate to crime(s) against humanity, a phrase which is also found in politics and history, though less frequently. In contrast, philosophy is concerned with humanity in general, with our ‘common humanity’ and how ‘humanity can survive’ together, accounting for 12 of the 28 instances of ‘humanity’ in the philosophy texts where it has a wider range of uses, including concern for the survival/the fall/the future of humanity. In English it is used in relation to specific individuals and their
relation to acts of human kindness, as in the example of Mr Knightley, while in classics it
tends to have more negative associations around the loss of humanity, as shown in Table 3.2.

Although the numbers of instances are small, such disciplinary variations in form, meaning and use are characteristic of academic language. In contrast with the BNC, expressions such as ‘dregs of humanity’ are not found.

**COCA**

A search in the BYU Corpus of Contemporary American English (COCA) includes five ‘genres’: spoken, fiction, popular magazine, newspaper and academic, evenly balanced for size over five-year spans from 1990 to 2015. Here ‘humanity’ is used least (10.13 pm) in spoken texts and most (34.85 pm) in academic journals. Among the academic disciplines, ‘humanity’ is used extensively in philosophy and religious studies (179.37 pm), an average amount in humanities (35.05 pm), and perhaps surprisingly it is seldom used in education (8.05 pm) and medicine (1.79 pm). Further investigation might explore whether these are consistent characteristics of the disciplines or accidental findings reflecting topics in the articles selected for the corpus.

The top collocates include *against, crime(s), common and mass*. Interestingly, the second most frequent collocate is *habitat*, which introduces ‘Habitat for Humanity’, which occurs in newspapers rather than in academic texts.

Give just $35 to [Habitat for Humanity](https://habitat.org/haiti) to help repair one damaged home in Port-au-Prince

here’s a great idea: Volunteer with [Habitat for Humanity](https://habitat.org)

A good way of recycling cabinets and appliances is to donate them to [Habitat for Humanity](https://habitat.org) or repurpose them . . .

This phrase is also found in one news item in the BNC, but, unsurprisingly, not in the BAWE. ‘Humanity’ occurs throughout the decades in COCA, and most frequently between 2005 and 2009, where it is associated with *crimes against humanity*, such as genocide or violence against innocent civilians. The examples point to this being a period of news reporting and academic discussion of crimes against humanity in countries such as Sudan, Zimbabwe, the former Yugoslavia, Angola and Iraq.

**Table 3.2 Concordance lines of ‘humanity’ in the discipline of classics in the BAWE corpus**

<table>
<thead>
<tr>
<th>Classics</th>
<th>As the play goes on we watch as Medea’s humanity falls away.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classics</td>
<td>effectively shutting off the last of her humanity as, when</td>
</tr>
<tr>
<td>Classics</td>
<td>argue against the games, but not for the case of humanity. Instead, these members</td>
</tr>
<tr>
<td>Classics</td>
<td>really a lady who does not possess any humanity Representing</td>
</tr>
<tr>
<td>Classics</td>
<td>in the conventional sense, involves loss of humanity ‘. The final sacrifice</td>
</tr>
<tr>
<td>Classics</td>
<td>good of the future Rome, and so can submit to his humanity . This is not to say however,</td>
</tr>
<tr>
<td>Classics</td>
<td>of Caesar, although he tries to diminish his humanity Caesar had</td>
</tr>
</tbody>
</table>
Further searches show that *humanity* in COCA has four main uses: crimes against humanity (see *crimes, crime, against* and *genocide* in Figure 3.1), Habitat for Humanity (3.87 percent of the instances of *habitat* occur with *humanity* with 7.89 mutual information (MI), which indicates the strength of the collocation), *humanity* in contrast with God (*God* accounts for 0.22 percent with 3.77 MI), and *humanity* meaning compassion, while *humanities* in COCA is associated with features of American university culture, such as *college, endowment* and *majors*.

**COHA**

A search in The Corpus of Historical American English (COHA) suggests that *humanities* is most frequent in the latter part of the twentieth century, occurring (Figure 3.2) more than 80 times in the 1960s–1990s, compared to fewer than 20 times in the 1820s–1890s.
Earlier examples, as this one from the 1840s, are invoked in relation to education:

this process might be called humanization, i.e., the complete drawing out and unfolding of his proper nature making him perfectly a man-realizing his ideal character; (and hence, with singularly beautiful appropriateness, the proper studies of a liberal education used to be called, not only the Arts, but the Humanities).

The change in meaning referred to is also reflected in the relatively strong and enduring collocate ‘so-called’ which suggests humanities is not an established term (Table 3.3).

The strongest collocates at the end of the twentieth century (science, arts, social, national, endowment) are similar in many respects to those in Figure 3.3, reflecting a fully institutionalised meaning of humanities alongside arts and social sciences with departments, professors, and endowments. No collocates are found, however, of digital or research, suggesting these came later.

Hansard

Hansard is a transcribed record of speeches and debate in the UK House of Commons and House of Lords (https://hansard.parliament.uk/about, accessed 17/04/2018). A search for the term ‘humanities’ in the Hansard corpus points to the debate over a Humanities Research Council in the 1990s where we see the humanities decried as a ‘waste of time’ and where ‘so-called humanities’ appears pejoratively.

GloWbE

The various digital humanities movements perhaps aim to establish a clear, contemporary identity for humanities, and a search in the 1.9-billion-word GloWbE corpus of web material from 20 English-speaking countries in 2012–2013 reveals that the term digital humanities...
appears on the Web extensively in the UK, Canada, Australia and New Zealand, but to a lesser extent in India, Malaysia and Pakistan, and not at all in Hong Kong, Jamaica or African countries such as Ghana, Kenya, Tanzania and South Africa (Figure 3.3).

The one GloWbE instance of ‘so-called humanities’ is from an American site, but refers to Europe:

The European educational system’s current process of so-called reform is marked by de-financing, cuts, job losses, and also by a downsizing of non-rentable disciplinary fields – the so-called humanities – accompanied by the increased support of capital-intensive fields of research. The leading principle of this reform is the assertion of the epistemological primacy of the economic sphere.

These searches for humanities show that while corpus investigations can produce statistical information about frequencies and collocations, it is important that users pay heed to the nature of the data in the corpus to inform interpretations. An understanding of the nature of the texts in the corpus, the contexts of their production and how they were selected all play a role in interpreting the significance of findings. We have glimpsed how pre-loaded corpora can be used to demonstrate the range of grammatical patterns in large general corpora associated with search items (through WordSketches), to suggest lines of enquiry (e.g. to what extent is humanities a contested term?) and to compare data from comparable sources (e.g. from different national web domains or from student writing in law vs. literature). To answer specific research questions using written corpora, however, requires careful attention to corpus selection, preparation and interrogation.

**Small correspondence corpora**

This study uses three small correspondence corpora to explore some of the linguistic features that are typically found in letters. The **Migrant Correspondence Corpus** (MCC) contains 188 private letters (176,501 tokens) between Irish migrants and their families, dating from 1819 to 1953; roughly 75 percent of the letters are by male authors (141 letters) and 25 percent (47 letters) by female authors. The **British Telecom Correspondence Corpus** (BTCC) contains 612 business letters (130,000 tokens) dating from 1853 to 1982. The letters – sourced from the public archives of BT – are predominately written by men, with only 13 authors identifying themselves as female. Finally, the **Letters of Artisans and the Labouring Poor** (LALP) corpus contains 272 letters (51,408 tokens) dating from 1720 to 1837 written by individuals appealing for poor relief from their parish or poor law authority; 118 male and 66 female authors are represented in the corpus. For this study letters from Record Offices in Warwickshire, Lancashire and Essex were used. The corpora represent different types of correspondence (personal and professional letters, as well as letters of solicitation), allowing users to explore the discourse of correspondence from a diachronic perspective. This analysis will investigate the register of correspondence, examining the extent to which the type of letter (personal or professional, for instance) predicts variation in the language. It will also explore how the MCC, BTCC and LALP data compare to modern English language registers of other kinds.4

The MCC, BTCC and LALP letters are all saved in XML – a format that is compatible with most corpus software. Each collection has an accompanying spreadsheet containing metadata relating to the individual XML letter files. Metadata provides additional information or knowledge about the text. It can be situated within the body of the text, providing
information or knowledge relating to the structure, layout or content (line breaks, paragraphs or pragmatic features, for instance); and it can be situated outside the body of the text – in the header – providing information or knowledge about where the text comes from (bibliographic information) and what it is (a letter, an academic essay or a poem, for instance). Most corpora will contain some basic header information; however, the type of information captured will vary from project to project depending on the research aims. While the MCC header information includes extensive sociobiographic details about the various participants (their sex, occupation, educational background and family history), BTCC header information focuses more on the professional status of the various authors and recipients and the communicative function of the letter itself. The LALP corpus, in contrast, captures header information about the materiality of the letter – whether it was written in pen or pencil and the quality of the paper, as well as details about the authenticity and authorship of the correspondence (whether or not the letter was dictated, for example).

Spreadsheets and databases are ideal for capturing well-structured metadata such as the header information described earlier. However, to use this metadata with corpus tools, it must be encoded in a way that the software will recognise. While there are many ways of encoding metadata, the Text Encoding Initiative (TEI) is the de facto standard for encoding digitised texts in the humanities. Although many of the information categories used in the MCC, BTCC and LALP corpus are varied, they do share some common features: a sender, a recipient, an origin, a destination and a date. The TEI has recently introduced a module for capturing this information in the <profileDesc> section of the TEI header (see Figure 3.4), allowing different letter collections to interconnect based on these shared attributes (see Stadler et al. 2016). Provided the information categories are TEI compatible, it is relatively straightforward from a programming perspective for metadata to be extracted from the spreadsheet and rendered as TEI headers within the corresponding XML files. These XML files can then be uploaded into corpus software, allowing the user to refine their search queries based on the header information (searching for all letters from a particular period or location, or by a particular author, for instance).

```xml
<profileDesc>
  <correspDesc xmlns="http://wiki.tei-c.org/index.php/SIG:Correspondence/
task-force-correspDesc">  
    <correspAction type="sending">  
      <persName>Elizabeth Lough</persName>
      <settlement>Winsted, Connecticut</settlement>
      <date when="1876-03-07"/>
    </correspAction>
    <correspAction type="receiving">  
      <persName>Elizabeth McDonald Lough</persName>
      <settlement>Meelick, Queen’s County</settlement>
    </correspAction>
  </correspDesc>
</profileDesc>
```

*Figure 3.4* TEI encoding for capturing information about the sender, recipient, origin, destination and date, using metadata from the MCC as an example.
To get a sense of how the letters in the MCC, BTCC and LALP corpus compare with other English language registers, the data were uploaded into Nini’s (2015) Multidimensional Analysis Tagger (MAT). MAT draws on Biber (1988) by grammatically annotating the corpus and plotting the distribution of corpus holdings based on Biber’s (1988/1995) dimensions. Dimension 1, ‘Informational versus Involved Production’, represents ‘a fundamental parameter of variation among texts in English’ (Biber 1988: 115). While one end of the pole represents ‘discourse with interaction, affective, involved purposes’ (telephone or face-to-face conversations, for example), the other end represents ‘discourse with highly informational purposes’ (official documents, or academic prose etc.) (ibid.). In terms of correspondence, Biber’s study showed that personal letters had a mean score of 19.5 for dimension 1 (placing them alongside spontaneous speeches and interviews), and professional letters had a mean score of –3.9 (placing them alongside general fiction and broadcasts). In other words, although both types of correspondence are interactional (as evidenced by the high frequency of first- and second-person pronouns), they differ in terms of their functions: while personal letters are more interpersonal, sharing many of the same characteristics as conversations, professional letters are more informational, sharing characteristics typically found in written discourse. Table 3.4 compares the MCC, BTCC and LALP corpus with Biber’s mean scores for personal letters and professional letters.

Similar to Biber’s findings for professional letters, Table 3.4 shows that the BTCC letters have a mean score of –7.1 for dimension 1, placing them at the ‘informational’ end of the ‘Informational versus Involved Production’ spectrum. However, unlike personal letters with a score of 19.5, the MCC letters have a mean score of 0.36, placing them alongside general fiction. In other words, the migrant letters are much less ‘affective’ and ‘involved’ than the personal letters used for Biber’s study. The LALP letters are quite interesting; with a mean score of 2.17, they are more conversation-like in nature than the MCC letters and correspond most closely with Biber’s prepared speeches.

Having looked at the dimension scores for the three corpora, the next stage involved examining some of the linguistic features that are common across all three datasets: namely, first- and second-person pronouns and what Biber describes as private verbs – that is, verbs which ‘express intellectual states or nonobservable intellectual acts’ (ibid.) – typically verbs of cognition such as ‘I think’ or ‘she hopes’. In systemic-functional grammar (e.g. Halliday and Matthiessen 2004) these clauses can function as a type of projection. Projection structures consist of two main components: the projecting clause (I hope) and the projected clause (you will write). In these structures the primary (projecting) clause sets up the

<table>
<thead>
<tr>
<th>Table 3.4 Mean scores for dimensions across five datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dimension 1</strong></td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td><strong>Dimension 2</strong></td>
</tr>
<tr>
<td><strong>Dimension 3</strong></td>
</tr>
<tr>
<td><strong>Dimension 4</strong></td>
</tr>
<tr>
<td><strong>Dimension 5</strong></td>
</tr>
<tr>
<td><strong>Dimension 6</strong></td>
</tr>
<tr>
<td><strong>Dimension 7</strong></td>
</tr>
</tbody>
</table>
secondary (projected) clause as the representation of the content of either what is thought or what is said (Halliday and Matthiessen 2004: 377). Halliday and Matthiessen make a distinction between the projection of propositions and the projection of proposals as follows:

Propositions, which are exchanges of information [typically statements or questions], are projected mentally by processes of cognition – thinking, knowing, understanding, wondering, etc. . . . Proposals, which are exchanges of goods-&-services [typically offers or commands], are projected mentally by processes of desire.

(2004: 461)

Both propositions and proposals have different response-expecting speech functions, and an analysis of these structures will reveal something about how the author interacts with their intended recipient and the type of response they expect – whether that is a verbal response (to provide information) or a non-verbal response (to carry out an action).

Sketch Engine (Kilgarriff and Kosem 2012) was used to automatically tag the corpus data for part of speech (POS), using the Penn Treebank tagset, making it possible to use Corpus Query Language (CQL) to identify lexico-grammatical patterns within each dataset. The following patterns were extracted (variants, e.g. with other pronouns or intervening adverbs, are not considered here.):

- [tag="PP"] to search for all personal pronouns
- [word="I"] [tag="V .."] to search for all personal pronoun + verb combinations
- [word="I"] [tag="V .."] [word="you"] to search for all personal pronoun + verb + personal pronoun combinations

Looking at the normalised frequencies (averaged per million), Table 3.5 shows similar frequencies for personal pronouns in the MCC and LALP (87,794 and 84,033, respectively); however, there are significantly fewer in the BTCC (51,113). It is a similar trend for the patterns I + Verb and I + Verb + You.

Having looked at the frequency of the pattern I + V + you, the next stage was to see which verbs typically appear in this structure (see Table 3.6). Not all of the examples in Table 3.6 are types of projection. I thank you, for instance, is a straightforward subject/verb/object construction. The most common verbs of desire (which typically realise the projection of proposals) are hope and wish, occurring across all three corpora. The verb want also occurs in the MCC and BTCC. The most common verbs of cognition (which typically realise the projection of propositions) are think (across all three corpora), suppose (in the MCC and BTCC) and know (which has a high frequency in the MCC, but only occurs once in the BTCC and not at all in the LALP corpus). Interestingly, the LALP corpus contains

<table>
<thead>
<tr>
<th></th>
<th>MCC</th>
<th>BTCC</th>
<th>LALP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr</td>
<td>16,814 (87,794 p/m)</td>
<td>7,511 (51,113 p/m)</td>
<td>4,332 (84,033 p/m)</td>
</tr>
<tr>
<td>I + V</td>
<td>3,951 (20,630 p/m)</td>
<td>1,821 (12,326 p/m)</td>
<td>912 (17,728 p/m)</td>
</tr>
<tr>
<td>I + V + you</td>
<td>212 (1,106 p/m)</td>
<td>73 (494 p/m)</td>
<td>48 (933.05 p/m)</td>
</tr>
</tbody>
</table>
a high frequency of verbs of desire (hope, wish, desire), but only one verb of cognition (think); additionally, there are several instances of what Biber (following Quirk et al. 1985: 1182–1183) describes as suasive verbs (verbs that ‘imply intentions to bring about some change in the future’ (1988: 242) – beg, solicit and ask).

I hope you is the most frequent projection structure across all three corpora; however, a closer examination of the language in context revealed some differences. In the MCC, I hope you is typically used in formulaic greetings (21 out of 37 occurrences – see concordance lines 1 and 2 in Figure 3.5, for instance). In the remaining 16 occurrences, the recipient of the letter (you) is typically required to think, pray, or remember/not forget (see concordance lines 3 to 6). In the BTCC only 3 out of 13 instances of I hope you are part of a formulaic greeting. In over half of the occurrences, what follows are the modals may, will or would + agree (see concordance lines 7 to 9). Finally, in the LALP corpus, 32 of the 39 occurrences of I hope you are followed by will. Here the recipient is typically required to send money (concordance line 10), speak or write to somebody on the author’s behalf (concordance line 11) or consider the author’s request (concordance line 12).

What is significant about projection structures is their ability to directly address and involve the recipient of the letter, assigning to them a role to play in the communicative event that is taking place. In the MCC, the subject of the projected clause – you (typically a close family member) – is required to respond cognitively in some way, by remembering a relative or by not feeling abandoned or forgotten. Similarly, in the BTCC data, for the most part, the subject of the projected clause (a business associate) is also required to respond cognitively – this time, by agreeing to something. In contrast, the subject of the projected clause in the LALP data (the poor law authority) is almost always required to carry out a

Table 3.6 Instances of I + Verb + You across the three corpora, organised by frequency

<table>
<thead>
<tr>
<th>MCC</th>
<th>BTCC</th>
<th>LALP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I hope you (61)</td>
<td>I thank you (15)</td>
</tr>
<tr>
<td>2</td>
<td>I tell you (16)</td>
<td>I hope you (13)</td>
</tr>
<tr>
<td>3</td>
<td>I wish you (13)</td>
<td>I told you (6)</td>
</tr>
<tr>
<td>4</td>
<td>I suppose you (12)</td>
<td>I think you (5)</td>
</tr>
<tr>
<td>5</td>
<td>I know you (12)</td>
<td>I sent you (5)</td>
</tr>
<tr>
<td>6</td>
<td>I want you (11)</td>
<td>I wish you (3)</td>
</tr>
<tr>
<td>7</td>
<td>I told you (8)</td>
<td>I suppose you (3)</td>
</tr>
<tr>
<td>8</td>
<td>I think you (8)</td>
<td>I send you (2)</td>
</tr>
<tr>
<td>9</td>
<td>I send you (6)</td>
<td>I see you (2)</td>
</tr>
<tr>
<td>10</td>
<td>I left you (6)</td>
<td>I wrote you (1)</td>
</tr>
<tr>
<td>11</td>
<td>I sent you (4)</td>
<td>I wanted you (1)</td>
</tr>
<tr>
<td>12</td>
<td>I wrote you (3)</td>
<td>I trust you (1)</td>
</tr>
<tr>
<td>13</td>
<td>I expect you (3)</td>
<td>I trouble you (1)</td>
</tr>
<tr>
<td>14</td>
<td>I assure you (3)</td>
<td>I thought you (1)</td>
</tr>
<tr>
<td>15</td>
<td>I write you (2)</td>
<td>I telegraphed you (1)</td>
</tr>
<tr>
<td>16</td>
<td>I thought you (2)</td>
<td>I reminded you (1)</td>
</tr>
<tr>
<td>17</td>
<td>I see you (2)</td>
<td>I promise you (1)</td>
</tr>
<tr>
<td>18</td>
<td>I saw you (2)</td>
<td>I offer you (1)</td>
</tr>
<tr>
<td>19</td>
<td>I remain you (2)</td>
<td>I observe you (1)</td>
</tr>
<tr>
<td>20</td>
<td>I refer you (2)</td>
<td>I know you (1)</td>
</tr>
</tbody>
</table>
physical action – to write or send money, or to consider carrying out these actions. It would seem that certain projection structures (for example, *I hope you*) may be indicative of the letter writing genre; however, their use varies depending on the type of letter it is (personal, professional, solicitation). Additionally, the author–recipient relationship would certainly influence language choice. Moreton (2015), for instance, examining a collection of migrant correspondence, found that letters to siblings, nieces or nephews (i.e. an ‘inferior’ within the notional familial hierarchy) contained more projections of proposals (often realising indirect commands), while letters to parents (i.e. a generational ‘superior’) contained more projections of propositions (typically statements/exchanges of information). Similarly, Sairio’s study of letters by Samuel Johnson to two of his correspondents (Mrs Thrale, a close friend, and Lucy Porter, Johnson’s step-daughter) found that the use of first- and second-person pronouns as well as evidential verbs (such as *know*, *think* and *believe*) are ‘a relevant indicator of the closeness of the relationship’ (2005: 33): the closer the relationship, the more likely it is that these linguistic devices will be used.

**Future directions**

Despite the exciting opportunities offered by the digital humanities with regard to digitising written resources, several constraints may hinder future research. Differing laws and attitudes relating to copyright and intellectual property across disciplines, cultures and countries affect accessibility of resources and are among the barriers to creating and interconnecting corpora. Certainly, when working with historical manuscripts, the only way around the problem, as argued by Honkapohja et al., is to ‘work with repositories and persuade them to either digitise the manuscript material and to publish them under an open-access licence, or to allow scholars to photograph manuscript material themselves’ (2009: 10). Sustainability of digital resources is also an issue, which Millett argues is exacerbated, in part, by ‘the tendency of large electronic projects . . . to use complex custom-built software, which makes them particularly difficult to update’ (2013: 46). And, finally, undocumented and/or differing transcription and encoding practices can lead to the duplication of work and make data interchange more challenging. However, alliances between disciplines are starting to form, largely driven by the need for reliable digital transcriptions of manuscripts that can be used across a range of disciplines. The *Digital Editions for Corpus Linguistics* (DECL) project, for example, ‘aims to create a framework for producing online editions of historical manuscripts suitable for both corpus linguistic
and historical research’ (Honkapohja et al. 2009: 451). Ultimately, what is needed is a cross-disciplinary, cross-cultural and cross-institutional approach to digitisation and annotation of corpus data.

Notes


2 University of South Carolina (2011) Victorian lives and letters consortium. Available at: http://tundra.csd.sc.edu/vllc/.

3 A term coined by the Spanish writer Miguel de Unamuno in 1985. ‘It refers to the value of the humble and anonymous lives experienced by ordinary men and women in everyday contexts which form the essence of normal social interactions, as opposed to the lives of leaders and famous people that are generally accounted for in canonical histories’ (Amador-Moreno et al. 2016).

4 The letters used to create the MCC are from Professor Kerby Miller’s archive of Irish migrant correspondence, housed at the University of Missouri. Miller has published extensively on the topic of Irish migration, see, for e.g. Miller 1985; Miller et al. 2003. The authors would like to thank Professor Miller for providing access to the letters. The BTCC data can be accessed from: https://btletters.wordpress.com. For more information about the LALP project see https://lalpcorpus.wordpress.com/.

Further reading

1 Creating and digitizing language corpora series (Volumes 1–3), published by Palgrave Macmillan. This is a useful source of information about corpus design and creation.


More information about written corpora and related publications can be found at:

- The Oxford Text Archive (OTA) http://ota.ox.ac.uk
- The Learner Corpora lists www.uclouvain.be/en-cecl-lcworld.html
- The Corpus Resource Database (CoRD) www.helsinki.fi/varieng/CoRD/
- The Sketch Engine site, www.Sketch Engine.co.uk
- The Brigham Young University site, http://corpus.byu.edu

The OTA houses dozens of corpora that can be requested by researchers in a range of formats. In contrast, the Sketch Engine and BYU sites can be used for researcher-compiled corpora, and both include dozens of pre-loaded corpora that can be easily analysed by novices.

The Sketch Engine site developed initially by the late Adam Kilgarriff includes the British National Corpus and the American National Corpus; the Brown Corpus, the Literary corpora of Early English Books Online and Project Gutenberg; web corpora such as Web 2008, 2012, 2013, uWAC (British web corpus) and Wikipedia; as well as more specialist corpora that will be of interest to researchers in the humanities, such as web corpora of African and Asian English (Araneum
Anglicum), the CHILDS corpus, the SiBOL/port newspaper corpus, a corpus of law reports, of international art (e-flux) and academic English (BAWE).

The BYU site developed by Mark Davies includes GloWbE, the NOW corpus and a Corpus of Online Registers of English (CORE). In addition to two corpora of British English (the BNC and Hansard) and one corpus of Canadian English, there are four corpora of American English, including COCA, COHA, Time Magazine and contemporary soap operas.

References


Written corpora


