Corpus linguistics as digital scholarship: Big data, rich data and uncharted data

Terttu Nevalainen, Carla Suhr and Irma Taavitsainen

University of Helsinki

1. Digital linguistics: approaches and developments

The past ten years have seen the rapid rise of Digital Humanities (DH), which currently subsumes a wide range of digital activities in various humanities disciplines, including linguistics and philology. One often-quoted definition of DH comes from the UCLA Digital Humanities Program, which states that:

> Digital Humanities interprets the cultural and social impact of new media and information technologies—the fundamental components of the new information age—as well as creates and applies these technologies to answer cultural, social, historical, and philological questions, both those traditionally conceived and those only enabled by new technologies.[1]

Edward Vanhouette (2013) traces various strands of DH back to the common denominator of Humanities Computing. Many series of publications were launched in this multidisciplinary field, which also linked linguistic research and computers. But as computers have become the standard tools of the trade, they tend to be replaced in publication titles by the more data- and technology-oriented label "digital". For example, the journal *Literary and Linguistic Computing* is now *Digital Scholarship in the Humanities*, the change of title "reflecting the huge changes that have taken place over recent years".[2]

Computers continue to be part of the title of the book series that publishes this volume, which was founded in 1988 with the title *Language and Computers*: *Studies in Practical Linguistics* and dedicated to "corpus linguistics and related areas". In 2016 the subtitle of the series was changed to *Studies in Digital Linguistics*. The series homepage updates its current agenda by saying that "a comprehensive digitization of our textual universe" calls for "a concerted research effort uniting linguistics and other disciplines involved in language-related research."[3]

In this interdisciplinary context we may ask whether the term "corpus linguistics" has by now outlived its usefulness. We would not be the first to ask this question. It was already raised by Jan Aarts in response to Nancy Belmore's query in the Corpora list twenty years ago in 1998. The point made by Aarts, and revisited by Antoinette Renouf in her contribution to this volume, was that it "is

[1] See http://www.digitalhumanities.ucla.edu/about/what-is.html (27 March 2017).
[2] Quoted from https://academic.oup.com/dsh (27 March 2017).
[3] Quoted from http://www.brill.com/products/series/language-and-computers (14 July 2017).

an odd discipline that is called by the name of its major research tool and data source". This line of thinking could equally well apply to digital humanities, where humanities research is based on the digital medium and technologies making use of this medium. But, as noted above, DH holds the promise to provide new answers to both traditional research questions and those that can only be broached by means of digital technologies. This is also the case with corpus linguistics.

As with DH at large, there is no need to abandon "corpus linguistics" as the name of a linguistic specialization that creates and studies structured machine-readable collections of texts and seeks to provide answers to research questions enabled by such data and methods. This does not mean that the skills, techniques and linguistic understanding associated with corpus-linguistic research would not be applicable to other fields; on the contrary, this is where corpus linguistics can and does feed into digital humanities in general and, in return, has an opportunity to contextualize, enrich and reassess its data sources and methodologies to better meet new research challenges. The big picture might look like the one shown in Figure 1, where corpus linguistics is visualized as part of digital linguistics and the larger multidisciplinary field of digital humanities. Needless to say, the boundaries between the three are permeable.

Moreover, Figure 1 does not aim to map the whole terrain of digital linguistics. Interdisciplinary fields such as natural language processing (NLP), human language technology, and computational linguistics would also be subsumed under this broad heading. In fact, digital linguistics consists of closely intertwined research interests: corpus linguistics, for example, largely owes its grammatical annotation tools to these related specializations. Defining their respective boundaries falls outside the scope of this chapter. Suffice it to say that in areas such as big-data applications their interests converge, and both technical and subject know-how, automated processing of language data and human validation of the results, are needed (for some examples, see 2.1).

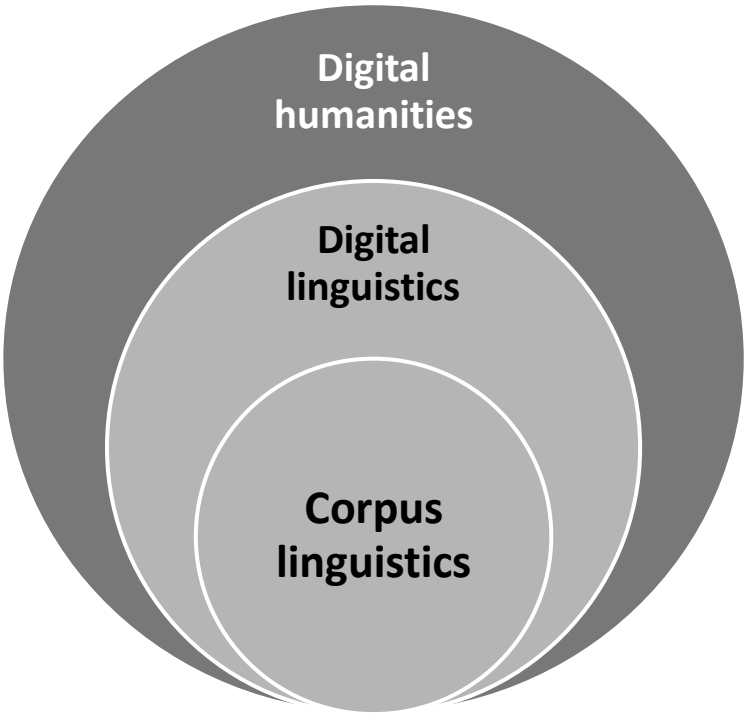


Figure 1. Corpus linguistics as digital scholarship.

As the title of this volume, *From Data to Evidence*, indicates, the contributions will be focused on the new affordances of text corpora and related data sources as well as of their digital processing, with a particular interest in the range of linguistic evidence they can generate. The aim is to show that corpus linguists are by no means at the mercy of their corpora – the question raised by Kytö and Rissanen in the context of early American English back in 1983 – but, informed by their subject expertise and benefiting from recent developments in DH, use their digital data sources in various innovative and creative ways to produce new linguistic evidence. This evidence will add to our understanding of the breadth and depth of language use, of linguistic constructions, and language variation and change, and will thus feed into linguistic theory, including usage-based modelling in linguistics. The role in theory building of linguistic evidence that accumulates from different data sources is the topic of, for example, Kepser and Reis (2005) and the subsequent Linguistic Evidence conferences. Their stated aim is to improve the empirical adequacy of linguistic theory and linguistic analysis by bringing together a large variety of data sources, including introspection, experimentation, language typology, and synchronic and diachronic corpora.[4]

Although the fascination of digital humanities lies in the potentiality to provide new answers to research questions, as cited above, "both those traditionally conceived and those only enabled by new technologies", the present paves the way for the future. In this volume our emphasis will be on

[4] See http://www.uni-tuebingen.de/forschung/forschungsschwerpunkte/sonderforschungsbereiche/sfb-833/ev/le2016.html (12 July 2017). Open-access publications and dedicated databases such as the Language Change Database (LCD) created and maintained in Helsinki for English historical linguistics naturally enhance the retrieval and processing of this evidence (http://www.helsinki.fi/lcd/; 14 July 2017).

the uses to which corpus linguists are putting the increasing variety of resources available for linguistic research. In concrete terms, we wish to bring into dialogue recent developments in data sources, tools and techniques, and linguists' creative and critical rethinking of how to apply them to meet their particular research needs.

These developments have been made possible, on the one hand, by a huge increase in computing power and the availability of techniques for retrieving, annotating and visualizing digital material. On the other hand, recent work is marked by a heightened awareness of what kinds of content is provided and what is left out from digital data sources that represent written, spoken and visual data as text. In this volume we distinguish three data-related processes that eventually lead to outcomes that can substantially enhance usage-based linguistic research:

(1) increasing the size of digital data sources – creating "big data";
(2) enriching context- and interaction-related information and developing tools – creating "rich data";
(3) discovering new data sources and rethinking existing ones – digging into "uncharted data".

These processes are naturally interconnected in practice. We would nevertheless argue that the resources resulting from them have certain distinct properties, both advantages and constraints, that set them apart from one another, depending on the linguistic uses made of them. Referring to the studies included in this volume, many of them based on historical data, we will demonstrate this point in the following sections and discuss the evidence on language and language use derived from various new or recent English-language corpora and databases and the ways in which they have been contextually and methodologically enriched for research purposes. Similar approaches can be, and have been adopted, in research into other languages. This is particularly the case with big newspaper databases and uses of the internet as a corpus (see 2.1).

Section 2 will introduce the three data-related processes outlined, pointing out some of the ways in which they are connected. By briefly introducing the individual chapters in this volume, Section 3 will discuss the kinds of linguistic evidence produced by these means in actual research practice.

2. Data-related processes

2. 1. Towards linguistic "big data"

Compiled in the 1980s and '90s, the one-hundred-million word British National Corpus (BNC) was the earlier benchmark for a very large corpus.[5] Today, linguistic big data cannot be defined in absolute terms for the simple reason that digital data sources are constantly being added to. Data may be collected to produce, for example, open-ended monitor corpora that grow on a daily basis. A case in point are newspaper corpora collected from online archives, such as the News on the Web (NOW), which at the time of writing covers some 4.7 billion running words of newspapers and magazines from 20 English-speaking countries from 2010 on, and is being augmented daily by millions of words, reaching an estimated total of 5 billion by the end of 2017.[6] The corpus is tagged and lemmatized, and standard corpus-linguistic tools such as concordancing and keyword searches are provided by the corpus interface. The metadata included make it possible for the corpus users to look up material from a particular date, country and newspaper, and by so doing create their own virtual corpora.

Digital data collections can also include multimodal information, which quickly makes the data to be analysed very large and complex indeed. The developers of the distributed Red Hen Lab write:

> While text corpora pose known problems and partial solutions, massive video corpora remain largely inaccessible to systematic analysis. Textual and visual information is complementary rather than duplicative, adding complexity to the parsing task.[7]

The Red Hen Lab is an example of a global consortium dedicated to the study of multimodal communication, which aims to develop a multilevel integrated research infrastructure with datasets from a variety of languages. The idea of joining forces with academic institutions also lies behind two large historical text creation partnerships, the Early English Books Online (EEBO-TCP) and the Eighteenth Century Collections Online (ECCO-TCP), which produce fully-searchable, SGML/XML-encoded texts of these massive text collections.[8]

The vast and still growing Google Books project benefits from library partnerships. Making use of this digital library, Michel *et al*. (2011) created the Google Books Ngram tool and launched the notion of "culturomics".[9] They provided quantitative analyses of the frequencies of a number of

---

[5] See http://www.natcorp.ox.ac.uk/corpus/creating.xml (21 July 2017).
[6] See http://corpus.byu.edu/now/ (16 July 2017).
[7] See http://www.redhenlab.org/home/the-cognitive-core-research-topics-in-red-hen/overview-research (21 July 2017).
[8] See http://www.textcreationpartnership.org/tcp-eebo/, http://www.textcreationpartnership.org/tcp-ecco/ (16 July 2017).
[9] See https://en.wikipedia.org/wiki/Google_Books; https://en.wikipedia.org/wiki/Culturomics (21 July 2017).

words between 1800 and 2000, arguing that "this approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology" (abstract). Figure 2 replicates one of their findings, the rising frequency of *women* as opposed to *men* in the late 20th century. The other, similar illustrations they present include names of scientists, popular dishes, and peaks of influenza epidemics.

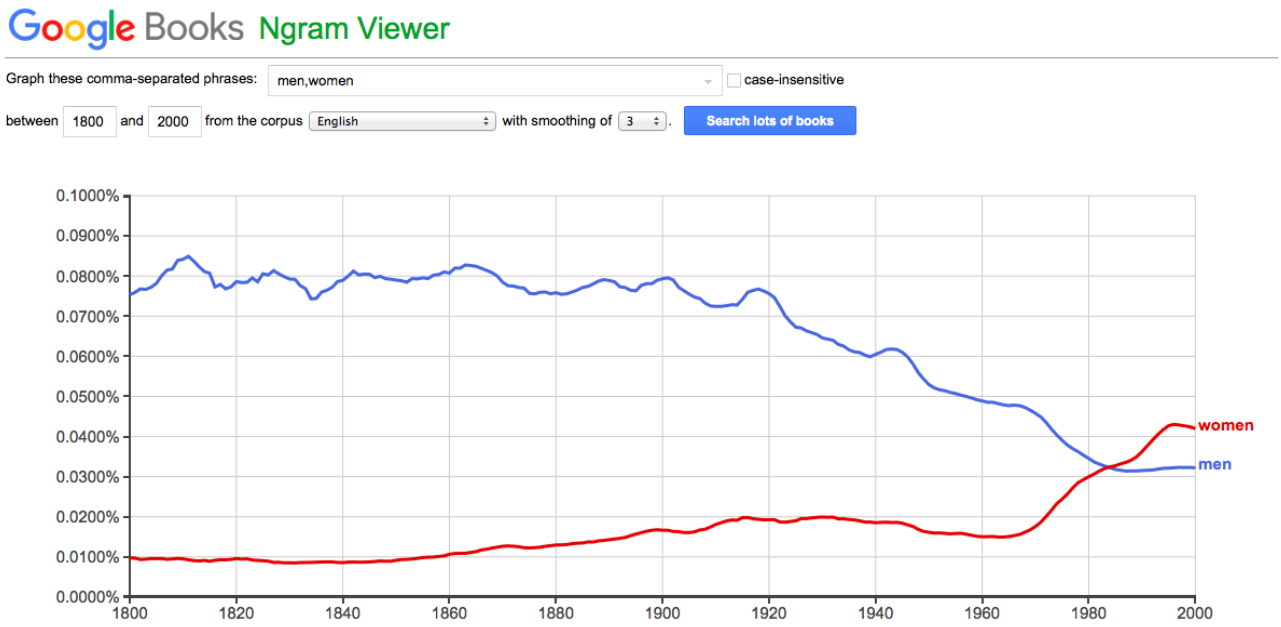


Figure 2. Frequency of occurrence of *men* and *women* (1800–2000) in the Google Books database.

In 2011 Mark Davies transferred the English and Spanish Google Books data into a format with more corpus linguistic functionalities than those offered by the original online interface, providing 155 billion words of American English (1.3 million books), 35 billion words of British English and 45 billion of Spanish.[10] The second release of the Google Books Ngram interface comes with more advanced search options, such as wildcards. Apart from American and British English, data are now available for Chinese, French, German, Hebrew, Italian, and Russian (versions from 2009 and 2012).[11] One of the disadvantages of this resource is that it does not provide any genre or register information, and hence cannot be used to produce linguistic evidence based on those variables or on the composition of the database (see e.g. Pechenick, Danforth & Dodds 2015, Koplenig 2017).

Taking the long view, there is a limit to corpus size, and the same yardstick cannot be used for historical corpora as is used for their present-day counterparts. For example, the entire Dictionary

[10] See http://googlebooks.byu.edu/ (19 July 2017).
[11] See https://books.google.com/ngrams (19 July 2017).

Old English Corpus (DOEC) consists of only three million words (see Table 1).[12] A corpus of the size of the DOEC would not by any measure count as a large corpus for Present-day English but it is all that has come down to us from the first six centuries of the language, c. 600–1150. It is obvious that the genre selection in the Old English Corpus would not be representative of the much wider range that has survived from later periods. But it does serve as the basis for *The Dictionary of Old English* (DOE), currently under compilation, and many of the same texts have provided the data for the Old English entries of the *Oxford English Dictionary* (OED)[13].

Table 1. The Dictionary of Old English Corpus (DOEC, 2009 release): text categories and word counts.

Category	Old English words	Foreign words
A: Poetry	177,480	255
B: Prose	2,128,781	52,038
C: Interlinear Glosses	699,606	635,655
D: Glossaries	26,598	70,511
E: Runic Inscriptions	346	4
F: Inscriptions in the Latin Alphabet	331	40
Total	3,033,142	758,503

Moving on to post-medieval times, corpora and databases grow larger. The EEBO database gives access to a vast repository of books published in English between 1475 and c. 1700. The resource comes close to qualifying as linguistic big data even according to current standards: the corpus version of the EEBO-TCP available on CQPweb server at Lancaster University contains over a billion running words (1,202,214,511, to be exact).[14] No detailed metadata is yet available for the EEBO-TCP database but, overall, religious content is highly prominent in print in the first half of the period. Moreover, as can be seen by comparing the breakdown of book titles in the EEBO-TCP sample of the database, the 17th-century part is much larger than the 16th-century one, let alone the 15th century.[15] The distribution of titles is directly reflected in the growing amount of text covered by the EEBO-TCP over time, as shown in Figure 3.

[12] See http://www.helsinki.fi/varieng/CoRD/corpora/DOEC/basic.html (16 July 2017).
[13] See https://en.wikipedia.org/wiki/Dictionary_of_Old_English, http://www.oed.com/ (16 July 2017).
[14] See https://cqpweb.lancs.ac.uk/, http://cass.lancs.ac.uk/?p=861 (27 March 2017).
[15] See http://www.textcreationpartnership.org/tcp-eebo/, https://earlyprint.wustl.edu/tooleeboestctexts.html (27 March 2017).

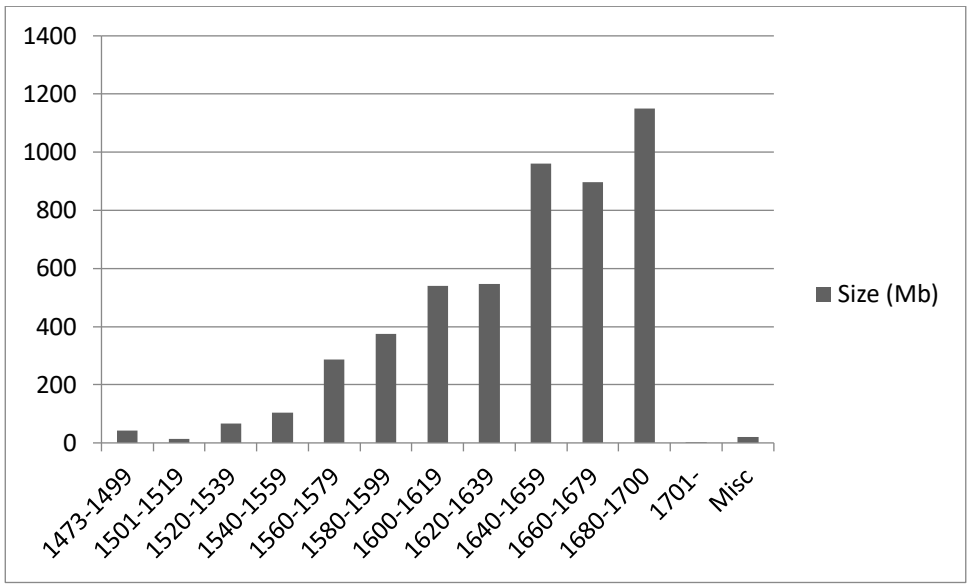


Figure 3. The size of the EEBO-TCP (first release).

As referred to in passing above, very large corpora and databases typically come with their online interfaces that provide the user with a range of search options and thus enable the sharing and reuse of these resources. The studies in this volume that are based on very large digital language data make use of the Brigham Young University (BYU) corpus architecture and interface,[16] the Corpus Query Processor (CQP) web-based corpus analysis system, which provides an interface to the Corpus Workbench (CWB) system,[17] and the WebCorp suite of tools, which gives access to the World Wide Web as a corpus.[18] All of them also provide material for the study of languages other than English.[19] Moreover, there are large integrated infrastructure projects, notably the Common Language Resources and Technology Infrastructure (CLARIN), an EU initiative, which has the mission "to create and maintain an infrastructure to support the sharing, use and sustainability of language data and tools for research in the humanities and social sciences".[20] Its consortium members currently include 19 European countries and two more with an observer status (Britain and France).

Corpus infrastructures based on large digital databases have also been created to serve specific purposes, to track neologisms, for example. Drawing on newspaper archives, the Logoscope tool detects and documents neologisms in French,[21] and the Néoveille platform aims to track lexical innovations in as many as seven languages: French, Greek, Polish, Czech, Brazilian

[16] See http://corpus.byu.edu/overview.asp (29 July 2017).
[17] See http://cwb.sourceforge.net/cqpweb.php (29 July 2017).
[18] See http://www.webcorp.org.uk/live/index.jsp (29 July 2017).
[19] Also worth mentioning is the Sketch Engine platform, which contains some 400 corpora in over 90 languages. See https://www.sketchengine.co.uk/ (29 July 2017). Sketch Engine charges a subscription fee.
[20] See https://www.clarin.eu/ (29 July 2017).
[21] See http://lilpa.unistra.fr/fdt/projets/projets-en-cours/logoscope/ (17 July 2017).

Portuguese, Chinese and Russian (Cartier 2016). The NeoCrawler project trawls the internet with the specific aim of identifying and detecting neologisms in English (Kerremans, Stegmayr & Schmid 2012).[22] The Monco search engines, in turn, provide live web-based corpora for several languages, which allow monitoring lexical innovations and their diffusion or, as the case may be, their failure to diffuse (as in the case of *Czechia* vs. *Czech Republic*).[23]

In general, the wealth of large digital resources opens up unprecedented research opportunities, including access to low-frequency lexical items and, ideally, these resources complement each other: newspapers generate neologisms of a different kind from those produced on Twitter, for example (for the latter, see Grieve, Nini & Sheng 2017). However, as will be discussed in section 3, linguistic big data can be problematic with respect to representativeness and comparability. Consisting of printed books, the Google Books database only reflects what has been published in that format. Professional and scientific publications seem to dominate this resource in the recent past, creating a skewed genre and register distribution (Pechenick, Danforth & Dodds 2015). The situation is not unlike that of early printed books, which was diachronically skewed by the dominance of religious material. Although this may be a true reflection of what has been printed, the corpus user should have access to the necessary metadata to be able to compare like with like in terms of genres and specialist domains over time.

As the size of digital resources reaches the proportions of billions of words, it is rarely possible for the data compiler or provider to supply them with the same level of descriptive metadata as is the case with smaller resources, typically "small and tidy" corpora (Mair 2006). The reasons for a lack of metadata may vary from the information simply not being available, which is often the case with historical data sources, to information not being collected for copyright reasons or because of privacy policies that control access to personally identifiable information. This is found with studies based on Twitter data, for example (Grieve, Nini & Sheng 2017). Rather than being the kinds of "precision tools" corpus linguists have been accustomed to, very large digital data sources become resources to be exploited for data exploration in various individual ways. The shift in the division of labour between the data provider and the data user therefore places an increasing responsibility on users to "know their corpus" (cf. Rissanen 1989).

2.2. Enriching context and developing tools

[22] See http://www.neocrawler.anglistik.uni-muenchen.de/crawler/html/ (17 July 2017). These infrastructures were among those discussed in Munich in June 2017 in a workshop dedicated to neologisms, http://www.anglistik.uni-muenchen.de/abteilungen/sprachwissenschaft/research/research_projects1/dfg-projekt/ws-lexinn/index.html (17 July 2017).

[23] See https://www.facebook.com/monco.en/ (17 July 2017).

The users' task of "knowing their corpus" can be aided by contextual knowledge provided by corpus compilers as metadata in the form of annotations or other supplementary information; this is what we call "rich data". Context is a multilayered notion that covers various text-external aspects from the micro surroundings of linguistic cotext to larger stretches of discourse, to groupings made on the basis of individual texts like genre and register to the all-encompassing cultural contexts that includes abstract notions like ideologies with political and religious commitments. The amount and type of contextual information that has become available for researcher in recent years, whether provided by corpus compilers or independent digital databases, have extended the range of potential research questions into areas such as sociolinguistics and pragmatics, where working with digital corpora needs to be complemented by background facts and related to the contemporary outside world.

The metadata can be encoded into the texts themselves, as relational databases (see Davies 2005), or as separate materials such as appendices or manuals. The information can be provided by corpus compilers or it can be added by corpus users themselves for their own research purposes. What is encoded varies a great deal in scope. The immediate cotext can be annotated automatically with parsers such as CLAWS or Penn Treebank for word classes and grammatical structure, though the labour-intensity of checking and completing the syntactic annotation has made corpora fully annotated in this way still quite rare.[24] The development of a program (VARD) for normalizing the spelling variation found in historical texts has extended automatic POS-tagging and grammatical parsing to these texts as well. For agglutinative languages such as Finnish, syntactic analysis requires morphological annotation as well.[25] Semantic tagging is also under development.[26] Metadata can also be provided about larger units of discourse (e.g. genre) and discourse participants. For example, both the *Parsed Corpus of Early English Correspondence* (PCEEC) and the Sociopragmatic corpus, a part of the *Corpus of English Dialogues* (CED), have been enriched with sociolinguistic speaker information, including parameters for sex, age and speaker role, for example (see Figure 3).[27] The Old Bailey corpus contains similar information about the participants in trial proceedings held at the Central Criminal Court of England and Wales.[28]

((METADATA (AUTHOR NICHOLAS_BACON_II:MALE:BROTHER:1543:26)
RECIPIENT NATHANIEL_BACON_I:MALE:BROTHER:1546?:23?)

[24] Examples of syntactically annotated corpora are the *Parsed Corpus of Early English Correspondence* (PCEEC), the Penn-Helsinki Parsed Corpus of Middle English, 2nd edition (PPCME2), and York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE).
[25] See the corpora and the Korp tool included in the Language Bank of Finland (https://www.kielipankki.fi/language-bank/, 26 September 2017).
[26] See http://ucrel.lancs.ac.uk/vard/about/, http://ucrel.lancs.ac.uk/annotation.html#acamri (27 March 2017).
[27] See http://www-users.york.ac.uk/~lang22/PCEEC-manual/corpus_description/index.htm (27 March 2017), http://www.helsinki.fi/varieng/CoRD/corpora/CED/index.html (27 March 2017).
[28] See http://www1.uni-giessen.de/oldbaileycorpus (27 March 2017)

```
        LETTER BACON_001:E1:1569:AUTOGRAPH:FAMILY_NUCLEAR))
(IP-MAT (CONJ nor)
      (NP-1 (D the) (N commyssion)
         (PP (P for)
            (NP (D the) (N pease))))
      (NP-SBJ (PRO I))
      (ADVP-TMP (ADV never))
      (VBD harde)
      (PP (P of)
        (NP *ICH*-1))
      (. .)) (ID BACON,I,7.001.5))
```

Figure 3. An annotated version for the extract "nor the commyssion for the pease I never harde of" in the PCEEC.

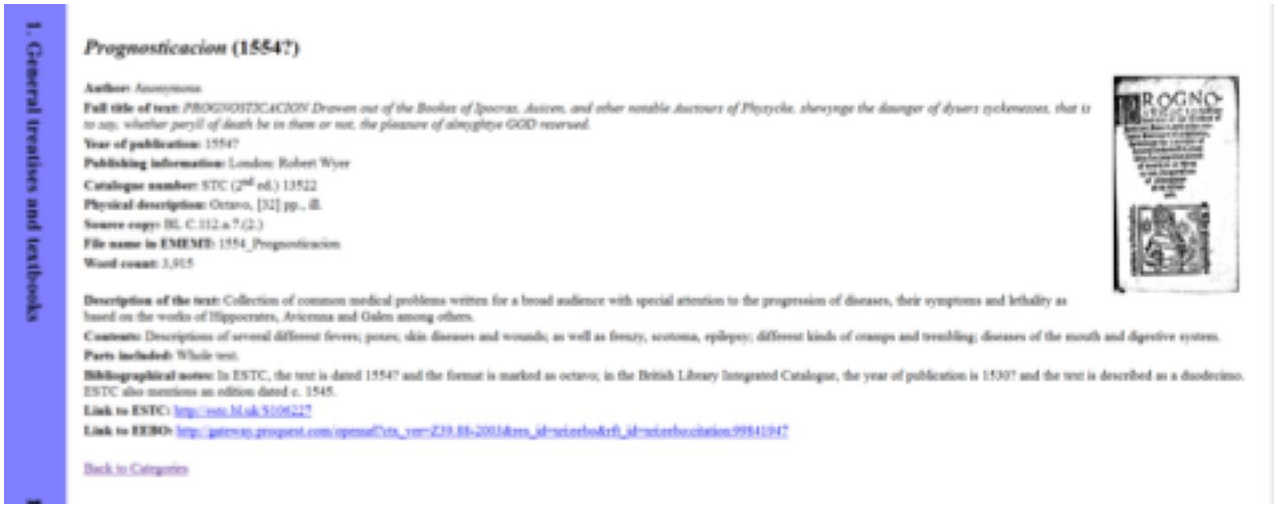
Pragmatic research questions profit from tagging of physical features of text, such as spacing, graphical elements and the choice of typeface in written data and, in spoken data, features such as prosody and gestures. For example, the *Middle English Grammar Corpus* (MEG-C) annotates, amongst other things, manuscript features such as abbreviations, flourishes, rubrics, underlinings and scribal corrections.[29] The move towards more detailed descriptive and analytic metadata has been aided particularly by the Text Encoding Initiative (TEI), which has provided an extensive annotation scheme for metadata that can be linked to either entire texts or individual textual elements.[30]

Contextual knowledge can also be provided separately from the corpus texts instead of as annotations. Manuals usually give general descriptions of the background and context of the corpus texts, but some corpora also provide extra materials that can help researchers analyse the corpus data. For example, the text catalogue of the *Corpus of Early Modern Medical Texts* gives the usual information on the author, publication history with details of use, if available, as well as descriptions of the physical book itself and its contents. Links are provided to external digital databases that can be consulted for further information. In addition, a picture gallery complements the corpus by providing the title pages of each text and the most important illustrations within them (see Figure 4). These images help in contextualising the books as objects of expensive or cheap production and they can also give indications of their use. Multi-modal corpora that include video and/or audio files in addition to transcripts as text files can also be considered rich data, as they

[29] See http://www.uis.no/research-and-phd-studies/research-areas/history-languages-and-literature/the-middle-english-scribal-texts-programme/meg-c/?s=8890 (27 March 2017) The Middle English Scribal Texts Programme at the University of Stavanger that produced MEG-C is currently working on *A Corpus of Middle English Local Documents* (MELD), which will also include extensive annotations.

[30] See http://www.tei-c.org/index.xml (27 March 2017).

provide contextual information such as gestures, facial expressions and prosody that may not be annotated into the text files but are nonetheless important for interpreting the situations correctly.


1. General treatises and textbooks

Prognosticacion (1554?)

Author: Anonymous

Full title of text: *PROGNOSTICACION Drawen out of the Bookes of Ipocras, Auicen, and other notable Auctours of Physycke, shewynge the daunger of dyuers sycknesses, that is to say, whether perill of death be in them or not, the pleasure of almyghtye GOD reserued.*

Year of publication: 1554?

Publishing information: London: Robert Wyer

Catalogue number: STC (2nd ed.) 13522

Physical description: Octavo, [32] pp., ill.

Source copy: BL C.112.a.7.(2.)

File name in EMEMT: 1554_Prognosticacion

Word count: 3,915

Description of the text: Collection of common medical problems written for a broad audience with special attention to the progression of diseases, their symptoms and lethality as based on the works of Hippocrates, Avicenna and Galen among others.

Contents: Descriptions of several different fevers; pestes; skin diseases and wounds; as well as frenzy, scotoma, epilepsy; different kinds of cramps and trembling; diseases of the mouth and digestive system.

Parts included: Whole text.

Bibliographical notes: In ESTC, the text is dated 1554? and the format is marked as octavo; in the British Library Integrated Catalogue, the year of publication is 1530? and the text is described as a duodecimo. ESTC also mentions an edition dated *c.* 1545.

Link to ESTC: http://estc.bl.uk/S106227

Link to EEBO: http://gateway.proquest.com/openurl?ctx_ver=Z39.88-2003&res_id=xri:eebo&rft_id=xri:eebo:citation:99841947

Back to Categories

Figure 4. An example of a catalogue entry in EMEMT.

The benefits of rich data are unquestionable, yet the use of rich data also presents challenges of its own. Archer (2012) discusses the difficulty of balancing between too much annotation and too general annotation: if the annotation scheme is too detailed, the less useful it will be for identifying general language patterns, but on the other hand, too general annotation schemes hide differences between text types. While corpus software that works with various kinds of file formats already exists, retrieval software that allows for annotation mark-up to be used as search terms is yet to be developed for many non-linguistic annotation systems. Adding annotations is usually labour-intensive if it can only be done manually. A case in point is speaker data that records, for example, discourse turns and speaker roles. Although micro-studies are valuable as such, a limited corpus size that is the result of intensive manual labour may diminish the generalizability of the research results. Limited corpus sizes can also prove to be problematic for data-driven studies that employ a variety of techniques that include statistics.

One way of solving or at least mitigating these problems is collaboration between corpus linguists and digital humanists in, for example, the fields of language technology and computer science in order to develop new tools and methods. However, not all research questions benefit from annotations, for some it is enough to have access to information that will allow the researcher to contextualize the results. A variety of tools exist in various digital databases to find the necessary contextual information. The ongoing wave of digitalisation has made available many databases that have their origins in printed resources. The point can be illustrated with the digital tools available for historical linguists. These include, for example, dictionaries such as OED, *The Middle English Dictionary* (MED), *The Historical Thesaurus of English* (HTE, now a part of the OED), *The Dictionary of National Biography* (DNB), linguistic atlases such as *The Linguistic Atlas of Early*

Middle English (LAEME) and *The Linguistic Atlas of Late Middle English* (LALME) and indices such as *The Digital Index of Middle English Verse* (DIMEV). *British History Online* is a digital library of primary and secondary sources. Hundreds of manuscripts have been digitised by the British Library and are freely available through their website (see the contribution to this volume McEnery & Baker). Primary sources are also available in subscription databases such as EEBO, ECCO, and *Nineteenth-Century Collections Online* (NCCO).[31] This list is by no means exhaustive, and only includes digital resources; many other resources for contextual information remain only in printed form or in manuscript repositories, and require researchers to familiarize themselves with earlier printed editions of texts and library archives.

2.3. Discovering new data sources and rethinking old

One of the most exciting prospects in the creation of new digital data sources involves the use of what we have labelled collectively as "uncharted data". The category comprises various kinds of material which has not yet been systematically mapped, surveyed or investigated. We wish to draw attention to the new research opportunities offered by texts and language varieties which are marginally represented in current corpora, to data sources that exist on the internet or in manuscript form alone, and to material compiled for purposes other than linguistic research. At the same time, existing corpora can be "recharted" or used in new ways by applying in their analysis new methods, either purpose-developed or imported from other fields of research. There is some overlap with our category of rich data here, as enriching existing data by adding metadata in the form of annotations could also be considered a method of rethinking old data sources.

The internet provides vast amounts of data that can be used to produce linguistic evidence. When the size of the data matters, we are dealing with big data (see 3.1). However, some big corpora allow their users to build their own smaller corpora from selections of the corpus material; NOW is an example of such a corpus that supports the compilation of "virtual corpora". Smaller corpora purpose-built for specific research designs can also be constructed from online material. Computer-mediated communication is a growing field that focuses on material generated online: e-mails, blogs, twitter feeds, chatrooms, discussion forums, just to name a few. In addition to providing new kinds of texts for linguistic analysis, the internet is a repository of older discourse forms such as news – and the range and scope of varieties of English made available online far exceeds that found in existing corpora of varieties of English. When compiling custom-built corpora

[31] For further information, see the following websites: www.oed.com; https://quod.lib.umich.edu/m/med; http://www.oxforddnb.com; http://www.lel.ed.ac.uk/ihd/laeme2/laeme2.html; http://www.lel.ed.ac.uk/ihd/elalme/elalme.html; http://www.dimev.net; http://www.british-history.ac.uk; https://eebo.chadwyck.com; https://quod.lib.umich.edu/e/ecco; http://www.gale.com/primary-sources/nineteenth-century-collections-online (27 March 2017)

from online sources, compilers face the same challenges of systematicity, representativeness and balance as other compilers, though their task may be made more difficult by too much data from which to choose rather than a dearth of data, as if often the case with historical corpora. For example, the variety of English, the accessibility of sources, the identifiability of authors and genre composition are just some of the key characteristics that Laitinen, Levin & Lakaw (in this volume) list as key components that need to be considered when compiling their multi-genre ELF corpora, collected from open sources.

In addition to material generated online, the internet also provides digital versions of existing texts. The digitalisation of materials is not often done for the purposes of linguistic study, but they can nonetheless be used as corpora if the users are aware of their limitations. EEBO is an example of a data source that has been turned into a corpus from a text repository. The searchable online edition of *The Old Bailey Proceedings*, the database called *Old Bailey Online*, has also been turned into the *Old Bailey Corpus*, with extensive enriching annotations added to the original texts. Digital editions of manuscripts can also be turned into corpora for linguists (see, for example, Marttila 2014, also available online).[32] New data, when they are small and custom-built corpora, also tend to be rich data.

Statistical methods such as cluster analysis and principal component analysis have long been used in corpus studies, but increasing contact with other digital humanists in neighbouring fields such as computational linguistics and information theory has exposed corpus linguists to new methods of analysis, which has in turn reverted existing corpora back into uncharted territory. These new methods can, on the one hand, be used to test in new ways existing hypotheses that are based on more traditional corpus-assisted analysis, but, on the other hand, they can also provide fresh research questions and novel insights that the statistical tools more familiar to corpus linguists simply cannot offer.

3. Linguistic evidence discussed in this volume

3.1. Evidence from "big data"

The contributions to this section all use very large corpora, the largest of them being the corpus of Global Web-based English (GloWbE 1.9 billion words), the Hansard Corpus (1.6 billion words), and the WebCorp Linguist's Search Engine diachronic corpus (WebCorp, c. 1.4 billion words). The other very large corpora discussed include the Corpus of Contemporary American English (COCA, c. 520 million words), the Corpus of Historical American English (COHA, c. 400 million words),

[32] See http://urn.fi/URN:ISBN:978-951-51-0060-3 (27 March 2017).

and the British National Corpus (BNC, c. 100 million words). In comparison with the earlier standard one-million-word corpora, these tried-and-tested, structured resources may be referred to as big data – or, as many contributors to this section prefer to call them, "very large corpora" – in English corpus linguistics, although corpus size is a moving target, and these linguistic resources would not qualify as such in many other data-rich disciplines. Much of the work on these very large corpora discussed in this section is theory-driven rather than purely data-driven (cf. Xiao 2008). In this sense, linguistic big data does not mean "the end of theory", a slogan often associated with big data analytics, which is theory-free in that its sole aim is to detect patterns and correlations of any kind (Hilbert 2016, 140).

The four chapters in this section all use very large corpora to explore lexis and lexico-grammar, providing evidence on innovative lexical developments (Renouf), on diachronic variation and change in lexis and semantics (Davies), changes in verbal syntax compared to prescriptions of normative grammar (Anderwald), and in alternative verb complementation patterns (Kaunisto and Rudanko). Although they do not necessarily use the corpora they refer to in their entirety, these empirical studies would not have been possible without access to very large structured corpora. Coming with a search interface and a corpus architecture that cater for lexico-grammatical studies in particular, corpora such as the Corpus of Historical American English also provide the researcher with access to a balanced structure of major genres over time. A similar structure was devised and implemented by David Lee (2001, 57–58) for the genres of the British National Corpus. Such convergent corpus structures naturally facilitate cross-corpus and cross-variety comparisons and generalizations based on them. Conversely, evidence from corpora with different structuring principles only allows more limited comparisons to be made.

What is of particular relevance in this volume is that the contributions also address problems to do with very large corpora, ranging from the degree to which these in fact meet the criteria set for linguistic corpora to issues of data granularity. **Antoinette Renouf** provides a critical assessment of both these issues in her chapter, which discusses the study of the rise of new words, lexical productivity and potential semantic change using very large newspaper corpora. Words in the medium frequency range normally pose the least problems for the corpus linguist, as she illustrates by the case study of *moot*. Renouf also shows the benefits of using a very large corpus for the analysis of low-frequency lexical items (typically unique occurrences, *hapax legomina*), which comprise over half of the word types in the corpus, but points out that there is no ready way to determine the extent to which they represent emerging usages rather than unintentional variation such as typographical errors.

At the other end of the frequency range, the analysis of high-frequency lexical words can become so unwieldy that it is no longer feasible to adhere to the principle of *total accountability*,

that is, using corpus data exhaustively, which has been one of the basic principles of traditional corpus linguistics. The collocational range simply becomes too diverse to manage. Renouf shows how these issues become of theoretical interest in lexicology and morphology, relating, for example, to derivational productivity, to rule blocking, and to detection of sub-word elements such as word-base categories. More sophisticated analytical software is called for to meet these challenges that are especially encountered by the lexicologists and lexicographers among corpus linguists, but which can also raise the question of research economy in socio-pragmatically oriented studies.

Mark Davies approaches similar issues by comparing the evidence provided by big and small corpora on the one hand, and big corpora and what he calls very large web-only corpora on the other. His focus is on lexical and semantic variation and the demands made on corpus size, for example, by collocational variation. The other major issue that he raises is the relevance of genre variation to lexical and syntactic phenomena ranging from adjective derivation to preposition stranding and the quotative *be like*. Comparing the distribution of these elements in large genre-aware corpora and a web-only corpus like GloWbE that does not make such distinctions shows that the latter resource gives very irregular results, and hence cannot be relied on as a source for the full range of lexico-grammatical variation in the language.

Davies offers three solutions to this problem. The first one involves creating a balanced sampling frame for a corpus to systematically record metadata such as dates, dialects, genres and authors etc., and storing this information in a relational database to allow for searches and cross-corpus comparisons of various kinds in a unified corpus architecture. The other alternative is to impose, for example, register structure on web-based texts post hoc, after the corpus has been collected. This was done for the CORE corpus (Corpus of Online Registers of English) using the Mechanical Turk, a crowdsourcing marketplace on the Internet, to assign register values to c. 50,000 texts.[33] The third option is to invite the corpus users to compile their own "virtual corpora" based on words within the texts or the titles of the texts, or various combinations of these.

Lieselotte Anderwald's study is concerned with innovative changes in verb syntax and morphology in the 19th century. Her aim is to trace any visible normative influence on a set of these processes over time. For this purpose, she compiled a large digital collection of 19th-century grammars (CNG) intended for native speakers of English and published in Britain and North America between 1800 and 1900. Using the genre-stratified data provided by the Corpus of Historical American English, Anderwald investigates the 19th-century trajectories of two constructions, the progressive passive (*the bridge is being built*) and the *get*-construction (*the house*

[33] See http://corpus.byu.edu/core/ (27 March 2017).

got built), and the past tense forms of two verbs, *leap* and *plead*. In each case she looks for observable peaks in prescriptive comments on these linguistic features in the grammar database prior to any major changes in their real-time trajectories in the four genres of COHA.

The results Anderwald obtained do not support any strong view on normative influence on actual linguistic practice in the 19th century. While most of the American comments on the progressive passive, for example, were highly negative, their impact on the diffusion of the construction only correlates with a temporary slowdown, mostly visible in newspapers. The other changes show even more modest correlations, or, as in the case of *plead*, the comments come after the verb form in question (*pled*) has gone out of use in these written sources. Putting her findings into perspective, Anderwald concludes that, relevant though it is, the corpus evidence we have for the potential impact of prescriptive grammars on language change is only part of the story and that prescriptivism has no doubt exerted a more lasting influence in social and psychological terms.

Mark Kaunisto and **Juhani Rudanko** are using several very large corpora to explore the extent to which a specific grammatical phenomenon is manifested in different varieties of English. They are interested in the use of the verb *warn* without a direct object or, in their terms, covert object control complement in the construction *warn against -ing* (*Mr. McCain will warn against making policy*), as opposed to the expected overt object control complement (*I would warn her against paying exorbitant prices*; the authors' examples). It is shown that the covert pattern is relatively recent, going back to the 20th century and attested in American English (COHA data) earlier than in British English (Hansard Corpus data), but the two patterns reached almost equal proportions in both varieties in the late 20th century. The recent British and American data in the GloWbE Corpus also show very similar proportions in the two patterns. The more limited material included in the corpus of Pakistani and Philippine English suggests that both these "outer circle" varieties are lagging behind the "inner circle" varieties in the diffusion of this change.

The fact that the patterns investigated fall within what Renouf calls the medium-to-low frequency band in a very large corpus makes it possible for Kaunisto and Rudanko to follow the principle of total accountability and examine all the relevant cases. They are also aware that the corpora they have used have different genre compositions, which makes them cautious in their generalizations of the results obtained.

The studies included in this section all suggest that one of the key issues in the use of very large data sets in corpus linguistics is the tools and infrastructure available to the researcher. If the users of very large corpora cannot always realistically aspire to the principle of total accountability, they should at least have the means to approach the issue in a principled manner. One solution, advocated by Davies for corpus-size comparisons, is replicating the findings obtained using other,

at least partly matching corpora.[34] But this clearly does not solve the issues arising from web-based unstructured big data, for example. This problem is shared by developers of big data resources in other fields of digital humanities as well. To quote the historian Tim Hitchcock (2014):

> In the rush towards 'Big Data' … the most urgent need seems to me to be to find the tools that allow us to do the job of close reading of all the small data that goes to make the bigger variety. […] This is not about ignoring the digital; but a call to remember the importance of the digital tools that allow us to think small; at the same time as we are generating tools to imagine big.

We will next discuss the ways in which this issue has been approached in concrete terms by those contributors to this volume who represent different linguistic specializations and have enriched their corpus-linguistic tools and resources accordingly.

3.2. Evidence from "rich data"?

The borderline between the categories of "rich" and "uncharted" data are fuzzy. In practice, nowadays new, uncharted data is often also rich data. Many of the chapters in these two sections move in both areas, and illuminate them from multiple angles. "Rich" can be translated to 'contextualised' and related to the text-external world, or it can mean 'enriched with annotation' to give short-cuts to the text-internal reality. The first two chapters included in this section for rich data deal with the latter definition by focusing on the pragmatic annotation of corpora (Kohnen, Rütten) to show how annotating metatextual information can help researchers, for example, identify relevant text passages, build textual networks, or recognize changing genre conventions. The last two chapters emphasise the contextualization reliant on corpus-external sources that is necessary for the initial stages of data selection and sorting in the pragmatic analysis of texts (Landert, Taavitsainen and Schneider). All four chapters have in common their data-driven approach and pragmatic research questions.

Thomas Kohnen begins this section with the theme of metadata annotation. The chapter enhances the potential of uncharted data and projects into the future by presenting a manifesto for metadata annotation of corpora. It speculates on what an ideal corpus of commonplace books would be like. The material is not even properly charted at present, and corpus compilation should start by mapping the "networks of multifunctional text reservoirs". Ideally the corpus should provide enriched entries of metadata for its users, as illustrated in the chapter. The books vary a great deal in

[34] For further discussion, see McEnery & Hardie (2011, 14 –16) and Nevalainen (in press).

their coverage, patterns of compilation and the repertoires of components vary. An annotation scheme of genre shifts should enrich the digital corpus as genre conventions differ from one another greatly. Besides e.g. availability of materials, the compositions reflect the linguistic practices of individual compilers of these notebooks. A corpus of commonplace books would open a window to the mindsets of their late medieval and early modern compilers and provide a welcome addition to the already existing digital corpora.

Tanja Rütten gives a practical example of the ways in which pragmatic annotation that details, for example, the genre, author, text user and network structure of a text would help research that considers larger textual structures and textual circulation. As her example she uses the prognostic texts included in the Dictionary of Old English Corpus (DOEC), though she points out that similar texts can also be found embedded in big present-day corpora such as GloWbE. The problem is that the texts cannot be identified easily, even in a small corpus such as DOEC, when there is no appropriate metadata annotation in the corpus; she uses external and contextual information to identify her prognostic texts in DOEC, but notes that such information is not available for big data. Rütten argues that small genres such as prognostic texts are hidden in larger corpora, which means that their coherent pragmatic and syntactic patterns also remain hidden in "the mass of the unfiltered output". She concludes that more precise and fine-grained metadata-annotation should be at the "top of the [philologist's] wish list".

Daniela Landert does not call for or rely on annotation in her study of stance markers in historical English, though part-of-speech tagging has often been used as an aid for identifying pre-selected forms of stance markers (see, for example, Biber 2004). Landert's aim is to chart comprehensively all the forms that stance marking takes in four register- or genre-specific corpora: the *Corpus of English Dialogues 1560–1760* (CED), the *Early Modern English Medical Texts Corpus* (EMEMT), the *Lampeter Corpus of Early Modern English Tracts* (LC) and the *Parsed Corpus of Early English Correspondence* (PCEEC). In order to do this, she has developed a method for automatically identifying text sections that are potentially rich in stance expressions; the method makes use of the fact that stance markers tend to cluster in texts. The text sections flagged for closer analysis revealed not only previously unstudied stance markers and potentially relevant contextual characteristics such as rhetorical questions, but the close contextual analysis of the sections also highlighted potential problems in quantifying the results. Of particular interest to other pragmaticians is the fact that Landert's method is scalable, so it can be used for data sets of different sizes, including big data, and that it can be used to identify other pragmatic functions in addition to stance markers.

The chapter by **Irma Taavitsainen** and **Gerold Schneider** also emphasises the importance of contextual knowledge to complement quantitative studies, at least when it comes to research

questions dealing with text structure and style. They employ a statistical tool new to corpus linguistics, Document Classification, to study scholastic text styles in three historical medical corpora covering the period 1375-1800,[35] complemented by a new, previously unknown Middle English text. The division of corpus texts into binary categories (scholastic vs. non-scholastic, early vs. late scholastic, Category 2 of MEMT vs. EMEMT) could only be done on the basis of solid contextual information about scholasticism and scholastic texts, but once the binary division was done the tool considered linguistic features in interaction with other features rather than in isolation to identify stylistic features that are distinctive to each class as well as diachronic developments. By combining their new quantitative method with close contextual analysis, Taavitsainen and Schneider demonstrate that scholastic argumentation patterns continued to be used in later periods, though with more critical overtones.

The chapter by Taavitsainen and Schneider straddles our categories of rich data and uncharted data. It has a heavy emphasis on contextual understanding garnered from both corpus-external and –internal information, which places it in our category of studies producing linguistic evidence from rich data, but it also employs a new methodology not previously used in historical linguistics, which we consider to be a way of rethinking old data that is comparable to finding uncharted data. This could also be said of Landert's chapter. In the following section, we show how the final five chapters have taken new approaches to existing data or compiled completely new data.

3.3. Evidence from uncharted data and rethinking old data?

Our definition of "uncharted" data refers to fresh data sources that are either created as completely new (Brett and Pinna, Degaetano-Ortlieb *et al*., Laitinen, Levin and Lakaw) or adapted to new uses (McEnery and Baker, Hiltunen and Tyrkkö), or newly (re)discovered old materials that have remained unknown to modern researchers (the chapter by Taavitsainen and Schneider in the previous section fulfills this criterion for part of their material). We also include in this category old data that is rendered uncharted because of a novel method of analysis (Degaetano-Ortlieb *et al.*). In practice, all of the material investigated in the chapters has become available only in the past few years, with the exception of the Wikipedia material used by Hiltunen and Tyrkkö.

Tony McEnery and **Helen Baker** make use of the texts of the seventeenth-century section of the EEBO database that have recently become available as a corpus of about one billion words. Their material could thus be characterised as uncharted big data. McEnery and Baker investigate the collocations of four terms, *beggar*, *rogue*, *vagabond* and *vagrant*, in order to determine how these

[35] The *Corpus of Middle English Medical Texts* (MEMT), the *Corpus of Early Modern English Medical Texts* (EMEMT) and the *Corpus of Late Modern English Medical Texts* (LMEMT) form the Corpus of Early English Medical Writing (CEEMW).

groups of criminalized poor were described in seventeenth-century England, and what kinds of attitudes writers displayed regarding these groups. The terms were carefully selected after reading parliamentary, administrative and legal documents available in the database *British History Online* and identifying frequently occurring terms. Their frequencies were also checked in the corpus. The analysis of the collocations is very much a qualitative analysis that relies on the textual context of the four terms as well as knowledge of the socio-cultural situation. In addition to describing the different ways in which the terms were used and the attitudes they reveal, McEnery and Baker also chart the diachronic developments of the terms during the course of the century. As a concrete result of their study they note that the corpus texts are now in the process of being sorted out into literary genres, which is a first step in the direction of the metadata annotation that Kohnen and Rütten call for in their chapters.

Stefania Degaetano-Ortlieb, Hannah Kermes, **Ashraf Khamis** and **Elke Teich** use as their corpus the recently released Royal Society Corpus (RSC), which contains some annotations (for example, part-of-speech, text type and author), making it both rich and uncharted data. They also use methodologies adopted from information sciences, entropy and surprisal, that are based on conditional probabilities of context (or in their case, cotext) rather than frequency-based measures. The aim of their study is, on the one hand, to test earlier results on the dense packing of information in scientific English by way of two case studies, and also to look for new, previously unidentified patterns by way of a third case study. They also chart the diachronic developments of the features they investigate (nominal compounding vs. prepositional phrases, modal verbs, and part-of-speech trigrams). This chapter is an example of linguistic evidence gathered by using methods on new data.

The chapter by **Turo Hiltunen** and **Jukka Tyrkkö** is a different kind of example of the use of uncharted data, as they make use of data that has existed for a while but that has been used for quite restricted purposes in linguistic research: Wikipedia articles. Their paper compares the use of academic vocabulary (analysed with the aid of AWL or the Academic Word List) in Wikipedia articles and research articles in three disciplines: economics, medicine and literary criticism. The Wikipedia material is a selection of texts from a large corpus, so their study is also an example of one way of dealing with the problem of big data by only using select parts of the vast data set. The statistical methods employed for the analysis, however, do not require close reading. The hierarchical cluster analysis and principal component analysis show that Wikipedia articles are quite similar to the research articles of the same discipline when it comes to their use of academic vocabulary; differences are primarily between disciplines rather than genres.

The final two chapters in this section introduce completely new and uncharted data. **David Brett** and **Antonio Pinna's** chapter deals with lyrics of popular songs, a genre that has largely been ignored until recently. The authors present a new corpus of ten million tokens based on an online

song archive that also contains considerable amount of metadata. The corpus is thus also an example of rich and small data. The corpus was gathered by web crawling the index pages of an online song repository using two pieces of software, and the material was divided into subgenres. Seven of them proved most important, and one of the main aims was to examine the lexico-grammatical differences between them. Their linguistic analysis focuses on lexical density and keywords. Preliminary results show that some keywords like “hip hop” and “heavy metal” were highly characteristic of their subgenres, while others like “pop” were less useful. Shared keywords suggest common thematic grounds for some subgenres but, on the whole, popular song lyrics is far from homogeneous.

Mikko Laitinen, Magnus Levin and **Alexander Lakaw** deal with the lingua franca use of English (ELF). The chapter describes two new multi-genre corpora of written language in which English is used as a second-language L2 resource, alongside with the native languages of Swedish and Finnish. They argue that new ELF corpora should be tailored for the genres that actually exist in the ELF setting and also include, for example, electronically-mediated communication. The corpus materials come from outside learned settings, and the corpora also contain a tweet component, which is very recent addition to data sources. The target sizes of the corpora render them small corpora. The chapter gives an account of the current state of the work and demonstrates the potentials with three cases studies on recent ongoing changes in English in comparison with available L1 corpus data. The first of the case studies gives an account of how the so-called subjective progressive is adopted in ELF, the second focuses on the modal system, and the third on typological profiling of ELF data. The comparison of ELF data with L1 data is a new vantage point that serves a broader purpose of illuminating grammatical variability on a broader basis and acknowledges the importance of new non-native varieties of English.

The chapters in this volume are evidence of the dynamism of the field of English digital linguistics in general and corpus linguistics in particular. Linguistic evidence is gathered from big data (or very large corpora), new uncharted and potentially rich and small data is collected or existing data is rethought with the help of new tools and analytical methods. At the same time, new methodologies are introduced to find new ways of both corroborating earlier research and to ask new kinds of research questions. This makes for very exciting times for corpus linguists and holds great prospects for digital scholarship!

References

Archer, Dawn. 2012. Corpus annotation: A welcome *addition* or an *interpretation too far*? In Jukka Tyrkkö, Matti Kilpiö, Terttu Nevalainen & Matti Rissanen (eds.), *Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources* (Studies in Variation, Contacts and Change in English, Vol. 10). Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/10/archer/

Biber, Douglas. 2004. Historical patterns for the grammatical marking of stance: A cross-register comparison. *Journal of Historical Pragmatics* 5(1). 107–36.

Cartier, Emmanuel. 2016. Néoveille, système de repérage et de suivi des néologismes en sept langues. *Neologica* 10. 101–131.

Davies, Mark. 2005. The advantage of using relational databases for large corpora: Speed, advanced queries and unlimited annotation. *International Journal of Corpus Linguistics* 10(3). 307-334.

Grieve, Jack, Andrea Nini & Dian Sheng. 2017. Analyzing lexical emergence in Modern American English online. *English Language and Linguistics* 21(1). 99–127. DOI: https://doi.org/10.1017/S1360674316000113.

Hilbert, Martin. 2016. Big Data for development: A review of promises and challenges. *Development Policy Review* 34(1). 135–174.

Hitchcock, Tim. 2014. Big data, small data and meaning. *Historyonics*. 9 November 2014. http://historyonics.blogspot.fi/2014/11/big-data-small-data-and-meaning_9.html. Accessed 29 March 2017.

Kepser, Stephan & Marga Reis (eds.). 2005. *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives* (Studies in Generative Grammar 85). Berlin: De Gruyter Mouton.

Kerremans, Daphné, Susanne Stegmayr & Hans-Jörg Schmid. 2012. The NeoCrawler: Identifying and retrieving neologisms from the internet and monitoring on-going change. In Kathryn Allan & Justyna A. Robinson (eds.), *Current Methods in Historical Semantics*, 59–96. Berlin: De Gruyter Mouton.

Koplenig, Alexander. 2017. The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets – Reconstructing the composition of the German corpus in times of WWII. *Digital Scholarship in the Humanities* 32(1). 169–188. DOI: https://doi.org/10.1093/llc/fqv037.

Kytö, Merja & Matti Rissanen. 1983. The syntactic study of early American English:The variationist at the mercy of his corpus? *Neuphilologische Mitteilungen* 84 (4). 470–490.

Lee, David Y. W. 2001. Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology* 5(3). 37–72.

Mair, Christian. 2006. Tracking ongoing grammatical change and recent diversification in present-day standard English: The complementary role of small and large corpora. In Antoinette Renouf & Andrew Kehoe (eds.), *The Changing Face of Corpus Linguistics*, 355–376. Amsterdam: Rodopi.

Marttila, Ville. 2014. Creating Digital Editions for Corpus Linguistics: The case of Potage Dyvers, a family of six Middle English recipe collections. PhD thesis, University of Helsinki. Available at https://helda.helsinki.fi/handle/10138/135589.

McEnery, Tony & Andrew Hardie. 2011. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak & Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014). 176–182.

Nevalainen, Terttu. In press. Using large recent corpora to study language change. In Richard D. Janda, Brian D. Joseph & Barbara S. Vance (eds.), *The Handbook of Historical Linguistics*, Vol. 2. Malden, MA & Oxford, UK: Wiley-Blackwell.

Pechenick Eitan A., Christopher M. Danforth & Peter S. Dodds. 2015. Characterizing the Google Books Corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE* 10(10). e0137041. https://doi.org/10.1371/journal.pone.0137041.

Rissanen, Matti. 1989. Three problems connected with the use of diachronic corpora. *ICAME Journal* 13. 16–19.

Vanhouette, Edward. 2013. The gates of hell: History and definition of Digital / Humanities / Computing. In Melissa Terras, Julianne Nyhan & Edward Vanhoutte (eds.), *Defining Digital Humanities: A Reader*, 199–153. Farnham: Ashgate.

Xiao, Richard. 2008. Theory-driven corpus research: Using corpora to inform aspect theory. In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics: An International Handbook*, Vol. 2, 987–1008. Berlin: Mouton de Gruyter.