sciendo

ICAME Journal, Volume 43, 2019, DOI: 10.2478/icame-2019-0004

Early Modern Multiloquent Authors (EMMA): Designing a large-scale corpus of individuals' languages

Peter Petré[1], Lynn Anthonissen[1,2], Sara Budts[1], Enrique Manjavacas[1], Emma-Louise Silva[1], William Standing[1] and Odile A.O. Strik[1]
University of Antwerp[1], Ludwig Maximilian University of Munich[2]

Abstract

The present article provides a detailed description of the corpus of Early Modern Multiloquent Authors (EMMA), as well as two small case studies that illustrate its benefits. As a large-scale specialized corpus, EMMA tries to strike the right balance between big data and sociolinguistic coverage. It comprises the writings of 50 carefully selected authors across five generations, mostly taken from the 17th-century London society. EMMA enables the study of language as both a social and cognitive phenomenon and allows us to explore the interaction between the individual and aggregate levels.

The first part of the article is a detailed description of EMMA's first release as well as the sociolinguistic and methodological principles that underlie its design and compilation. We cover the conceptual decisions and practical implementations at various stages of the compilation process: from text-markup, encoding and data preprocessing to metadata enrichment and verification.

In the second part, we present two small case studies to illustrate how rich contextualization can guide the interpretation of quantitative corpus-linguistic findings. The first case study compares the past tense formation of strong verbs in writers without access to higher education to that of writers with an extensive training in Latin. The second case study relates s/th-variation in the language of a single writer, Margaret Cavendish, to major shifts in her personal life.

1 Introduction

The English language is blessed with an unusually rich array of language corpora. A key objective of corpus compilation is to be able to examine language in a naturalistic setting, under the assumption that empirical observation is key to the understanding of any naturally occurring phenomenon. Collecting naturalis-

tic data is far from trivial, however. Corpus linguists have often pointed out that representativeness is a problematic concept not just theoretically, but also in its implementation. Most corpora contain an overrepresentation of written data or more formal registers than the bulk of language use. A key characteristic of the empirical reality of language that has been less explicitly discussed in corpus research is the observation that language only exists through its users. Usage-based linguistics assumes that language users build up a grammar in their minds by combining general innate cognitive abilities with rich input from outside. These individual grammars emerge in the language acquisition stage but, to the extent that language can be seen as a complex adaptive system (Steels 2000; Beckner et al. 2009; Bybee 2010; Ellis 2011; Van de Velde 2014; Schmid forthcoming), it is also likely that language users continue to fine-tune their grammars beyond childhood, and across the lifespan. This individual dimension of grammar interacts with a socially defined one. Individuals create representations of what they think is the norm within a particular social context (various such norms may co-exist), and try to emulate this norm. Even so, as language is a very complex phenomenon, it is very likely that considerable differences in cognitive representation continue to exist between individual language users. While some of these are not much more than idiosyncrasies, some differences may also be recurrent across certain dimensions. Some of these dimensions, such as generation cohorts (Sankoff 2005), communities of practice (Eckert 2000: 34–41; Kopaczyk and Jucker 2013), or age grading (Wagner 2012), have been captured by sociolinguistics. Others may be a mixture of social and cognitive factors, such as the impact of education on the processing of syntactic structures (Dąbrowska and Street 2006; Dąbrowska 2015), or are presumably predominantly cognitive, such as the decrease in *d/t*-deletion in past tense endings over the lifespan observed by Guy and Boyd (1990). Obviously, these recurrent factors, which depend on groups of individuals, will determine in key respects the properties of language at the aggregate level of a speech community. To better understand how the underlying dynamics of language act and react to the intersubjective reality of language use, then, we should take into account this individual dimension to the fullest in corpus research no less than in, for instance, experimental research.

The purpose of this paper is to present a newly compiled corpus, *Early Modern Multiloquent Authors* (EMMA), and to show how this corpus specifically meets these needs. Existing corpora of Early Modern English are varied in scope and size. Specialized corpora are rigorously compiled and representative of specific language uses, but generally small. Examples include the corpora of Early Modern Correspondence (e.g. PCEEC; Nurmi et al. 2006), English Dialogues

(CED; cf. Kytö and Walker 2006), and Early Modern English Medical Texts (EMEMT; Taavitsainen et al. 2010). Well-established multi-purpose corpora, such as the PPCEME (Kroch, Santorini and Delfs 2004), the Helsinki Corpus (Rissanen et al. 1991) and ARCHER (cf. Yáñez-Bouza 2011), approach a high degree of balance but are also relatively small in size. Extensive digitalization projects such as Early English Books Online (EEBO) and Eighteenth Century Collections Online (ECCO) provide digitized editions of writings by all British authors between 1470 and 1800. Yet they are unstructured databases rather than real corpora.

The EMMA corpus fills a gap by being a large-scale specialized corpus that allows for the in-depth analysis of individuals' language use against the backdrop of community-level usage. EMMA is of course not the first corpus that is set up with individuals in mind. Within language acquisition, the CHILDES corpus is another great resource of individual data of parent-child interactions. However, when it comes to rich data from adult language users, existing resources show clear limitations. Spoken corpora sometimes include some sociological and sociolinguistic information on the participants, but such information is generally fairly limited, not rarely to conform to privacy and data protection requirements. Historical sociolinguists have compiled several excellent corpora based on ego-documents (mostly letters), for instance, the Parsed Corpus of Early English Correspondence (Nurmi et al. 2006) and its 18th-century extension CEECE (cf. Nevalainen et al. n.d.). These corpora do include rich metadata about their authors. The most important drawback of these corpora is that they are generally fairly restricted in size.

EMMA (*Early Modern Multiloquent Authors*) is a sample of 50 of the most prolific – or 'Multiloquent' – English writers born in the 17th century, the majority of them intellectuals belonging to the London society. The compilation of EMMA forms part of the ERC-funded research project *Mind-Bending Grammars*. The corpus is designed specifically for the quantitative study of syntactic change across the lifespan of individual language users from various perspectives, including cognitive dynamics of linguistic knowledge, historical sociolinguistics and intragenerational versus intergenerational change. Using the corpus, the project wants to investigate the extent to which innovation and change is possible across the lifespan in the domain of syntax. Major goals include (i) to fundamentally advance the debate on how different intragenerational change is from intergenerational change; (ii) to determine to what extent syntactic changes co-evolve; (iii) how social and cognitive factors interact. While compiled for syntactic research, the corpus lends itself well to all kinds of linguistic research that benefits from the individual perspective. The following sections will

explain EMMA's design features (Section 2), its formatting and compilation procedure (Section 3), and how it was enriched with metadata (Section 4), as well as illustrate the opportunities EMMA offers by means of two small case studies and an overview of the scholarship already based on EMMA (Section 5). Section 6 briefly describes plans for future improvements. Finally, EMMA is a corpus that has been released under a Creative Commons Attribution-ShareAlike 4.0 International License, and is freely available. Details on availability are provided in Section 7.

2 Overall structure and selection criteria

The EMMA corpus is a large-scale specialized corpus that comprises the writings of 50 carefully selected authors across five generations. At the individual level we looked for authors who met three primary criteria related to balance and representativeness. The ideal candidate would fulfil all of these, but in practice not many individuals were a perfect match. In general, we strove for an optimal balance between them. In discussing each criterion we explain what form this balance has assumed in the final corpus.

Criterion 1: The authors produced a large body of work comprising **at least 500,000 words.** We defined 'work' very broadly as all writings that have survived, ranging from personal letters over pamphlets to plays and scholarly treatises. The size of individual oeuvres was estimated on the basis of provisional word counts of all digitized texts in EEBO-TCP (Phase I and Phase II), ECCO-TCP and Evans-TCP (see also Section 3), as well as inferred from the number of pages (taking a conservative 250 words/page as guideline) of volumes included in ECCO. The average author in our final selection has an oeuvre of about 1.6 million words (disregarding one outlier, Richard Baxter, who alone has 10.5 million). A few of them did not actually reach the 500,000 word target, as the original estimates on which the author selection was based went down after the identification of duplicates, non-authorial material and foreign language passages. Because of this, John Crowne, Joseph Addison and Benjamin Hoadly ended up slightly below the target. In the case of John Harris, we decided to exclude (except for the preface) his *Lexicon technicum, or, An Universal English Dictionary of Arts and Sciences* (1704). While this is his most famous work, the fact that it is a dictionary marks it as an unsuitable outlier in comparison with his own writings as well as the writings of his peers. Its size also disrupts the even distribution of work, as it is several times larger in itself than the rest of his oeuvre. While a sample might still have been included, the work involved in transcription and identification of lemmas authored exclusively by him made us

decide against this, given the budgetary restrictions we had. The result is that John Harris's oeuvre is rather small (222,000 words). Another author's subcorpus, that of Samuel Clarke, is with 237,000 words in a similar situation. In this case, plenty more is available as dirty OCR, but budgetary limitations prevented us from correcting more. The expansion of these lesser represented authors is an important objective of a future release of EMMA.

Criterion 2: The authors' work is **relatively evenly distributed across a long career**. Assessments were based on the available texts in our sources, in combination with information from biographical and bibliographical resources (cf. Section 4). We consider an author's career to start, pragmatically, with the text with the earliest text date in our corpus and stop with the last text date. It is possible that some authors had their debut earlier and went on longer, but if these data are not in our corpus, this information is not taken into account. The average length of our authors' active careers is 38 years (sample standard deviation = 11.1 years). Only three authors have an active career of less than 20 years. These are Margaret Cavendish (15 years), George Swinnock (16), and Aphra Behn (19). Increase Mather was active the longest, with his debut on his 21st birthday, and his last work being published 63 years later, in the year he died, at the age of 84. Debut ages range between 18 (George Whitehead) and 34 (four authors), with an average of 26 (sample standard deviation = 4.2 years).

Criterion 3: The authors show a **demonstrable link to London society**. While this is still a fairly heterogeneous group, and London was becoming more fragmented in the course of the 17th century, London citizens have been shown to display a higher number of weak ties as compared to people outside the metropolitan, and to a certain degree a shared London identity may be assumed (Archer 2000, Nevalainen 2015). Thirty-seven authors spent a considerable amount of time in London, on average 54 percent of their lives (sample standard deviation = 20%). This average would be even higher if their youth is disregarded. Colley Cibber leads this group, as he only spent five years out of his long life (86 years) outside London. The remaining thirteen authors are not strongly connected socially or geographically to London. Seven of them spent most of their lives in smaller cities or towns in England: John Flavell (Dartmouth as well as other places in Devon), George Swinnock (Maidstone, Great Kimble), Henry More (Cambridge), Daniel Whitby (Salisbury), Thomas Pierce (Oxford and Salisbury), John Bunyan (Bedford), Peter Heylyn (various places in Oxfordshire and Hampshire). Jeremy Taylor was born in Cambridge, but was a cleric in Wales and Ireland for most of his life. Roger Boyle and Jonathan Swift are somewhat connected to London in that they lived there five and six years respectively, but both spent the majority of their lives in Ireland. Increase and

Cotton Mather lived (mostly) in New England (Boston). Finally, Margaret Cavendish, Duchess of Newcastle-upon-Tyne, spent a large part of her life in exile on the European continent because of the Civil War. Despite their weaker ties to the London society, these authors have been included in the corpus, as they may still serve as a control group when looking at the spread of linguistic changes through the London-based social networks of the time.

While the three selection criteria listed above pertain to the life and oeuvre of the writers individually, we also paid attention to the relations *between* the authors in our sample. Many of them exhibit social, political, and stylistic connections to other individuals in the selection. The connections are typically close-knit, i.e. dense and multiplex (Milroy and Milroy 1992: 5), involving multiple communities of practice. The largest community is probably that of religious leaders, such as Richard Baxter, Gilbert Burnet, and John Tillotson, who were in continuous debate about the desired direction of the English church, a heated topic closely intertwined with national politics ever since the Church of England separated from the Roman Catholic Church in 1534. Another large community was that of playwrights and literary authors, such as John Milton, John Dryden, Richard Steele, and Jonathan Swift. These two communities were in turn closely connected because they moved in similar social circles. Several of them for instance got to know each other at university, as most had an Oxbridge degree. Another obvious connection was the Court. Three of the playwrights in our corpus (William Davenant, John Dryden, and Colley Cibber) became members of the royal household when appointed as Poet Laureates. As many as ten religious authors in EMMA were at some point royal chaplains, and in a similar position. That the two groups talked to each other is for example evidenced in John Dryden referring to John Tillotson as his 'judicious and learned Friend' (in the preface to his *Religio laici,* 1682; cf. Rivers 2004). Smaller communities include that of the Quakers (within our corpus included are George Fox, founder, William Penn, who brought Quakerism to the US, and George Whitehead), or the Royal Society (cf. Gotti 2013; active members in EMMA include Henry More, Robert Boyle, John Tillotson, Gilbert Burnet, John Harris, and Samuel Clarke; people who attended meetings at some point include John Dryden, Margaret Cavendish, Nathaniel Crouch, and William Whiston; Cotton Mather was corresponding member). A more detailed visualization of the social network connections is provided in Section 4.3.

At the level of the author selection as a whole, we valued a distribution across different professions. Each generation includes two playwrights[1], four clerics, one historian, and one scientist (including a mathematician and a doctor). Table 1 gives an overview of the authors in the EMMA corpus, their profes-

sions and their respective word counts (EM represents a sample of EMMA, see below). Figure 1 visualizes the lifespans of the authors and their active careers.

Table 1: Authors in the EMMA corpus; the first letter of the ID denotes the generation to which the author belongs

ID	Author	Description	Word count	
			EMMA	EM
101	Heylyn, Peter (1599–1662)	churchman, author	3,712,572	350,793
102	Prynne, William (1600–1669)	lawyer, author, political figure	4,957,265	470,377
103	Davenant, Sir William (1606–1668)	playwright	504,413	339,677
104	Fuller, Thomas (1607–1661)	churchman, historian	2,652,292	275,026
105	Milton, John (1608–1674)	poet	729,624	307,695
106	Taylor, Jeremy (1613–1667)	cleric, author	3,132,105	303,512
107	More, Henry (1614–1687)	philosopher	1,867,798	523,626
109	Baxter, Richard (1615–1691)	church leader, poet, theologian	10,319,036	437,055
110	Owen, John (1616–1683)	church leader, theologian	4,350,175	419,860
111	L'Estrange, Roger (1616–1704)	pamphleteer, author, politician, Licenser of the Press	2,015,050	388,806
	Total generation 1		***34,240,330***	***3,816,427***
201	Boyle, Roger (1621–1679)	soldier, dramatist, politician	790,412	207,933
202	Pierce, Thomas (1622–1691)	churchman	978,491	280,524
204	Fox, George (1624–1691)	Quaker founder	1,018,398	327,434
205	Boyle, Robert (1627–1691)	natural philosopher, chemist, physicist, inventor	2,082,984	545,636
206	Swinnock, George (1627–1673)	churchman	946,926	302,282
207	Bunyan, John (1628–1688)	writer, preacher	1,330,929	326,086
208	Flavell, John (1630–1691)	clergyman, author	1,627,802	283,271
209	Tillotson, John (1630–1694)	Archbishop of Canterbury	507,557	257,053
210	Dryden, John (1631–1700)	poet, playwright, critic, translator	1,715,258	387,254
211	Cavendish, Margaret (1623–1673)	philosopher, poet, scientist, fiction-writer, playwright	1,393,983	229,557
215	Phillips, John (1631–1706)	translator, secretary to Milton	1,456,167	339,492
	Total generation 2		***13,848,907***	***3,486,522***
301	Stillingfleet, Edward (1635–1699)	theologian, scholar	2,974,637	396,347
302	Whitehead, George (1637–1724)	Quaker leader	1,284,629	462,586
303	Whitby, Daniel (1638–1726)	theologian, biblical commentator	1,925,091	589,336
305	Mather, Increase (1639–1723)	puritan minister, colonist	1,503,461	583,093
306	Sherlock, William (1641–1701)	church leader	2,076,365	305,775
307	Keach, Benjamin (1640–1704)	preacher	2,102,014	316,099
308	Crouch, Nathaniel (1640–1725)	printer, bookseller, historian	1,791,125	257,346
310	Behn, Aphra (1640–1689)	playwright, poet, translator, author, spy	1,039,596	262,050
311	Crowne, John (1641–1712)	dramatist	473,022	305,929
312	Burnet, Gilbert (1643–1715)	philosopher, historian, bishop	3,167,554	435,477
313	Salmon, William (1644–1713)	doctor	2,889,362	329,378
314	Penn, William (1644–1718)	Quaker, founder of Pennsylvania	1,478,837	325,747
	Total generation 3		***22,705,693***	***4,569,163***

401	D'Urfey, Thomas (1653–1723)	writer, poet	961,267	344,231
402	Wake, William (1657–1737)	Archbishop of Canterbury	1,143,686	269,423
403	Dennis, John (1657–1734)	playwright	672,818	373,283
404	Dunton, John (1659–1733)	bookseller, author, publisher	1,177,388	300,466
405	Defoe, Daniel (1660–1731)	author, journalist, spy	4,080,303	455,245
406	Mather, Cotton (1663–1728)	minister, author, pamphleteer	2,465,566	448,243
407	Harris, John (1666–1719)	writer, scientist, priest	219,963	219,963
408	Swift, Jonathan (1667–1745)	author, poet, satirist, pamphleteer, cleric	387,000	290,647
409	Whiston, William (1667–1752)	theologian, historian, mathematician	508,279	335,742
410	Ward, Edward 'Ned' (1667–1731)	satirist, publican	905,106	316,906
	Total generation 4		***12,521,376***	***3,354,149***
501	Cibber, Colley (1671–1757)	playwright, actor, manager, Poet Laureate	589,993	423,960
502	Steele, Richard (1672–1729)	writer, politician	541,503	255,384
503	Addison, Joseph (1672–1719)	essayist, poet, playwright, politician	487,207	257,840
504	Oldmixon, John (1673–1742)	historian, author	942,189	336,473
505	Clarke, Samuel (1675–1729)	philosopher, clergyman	229,619	229,619
506	Hoadly, Benjamin (1676–1761)	clergyman, bishop	425,529	328,077
508	Jacob, Giles (1686–1744)	author, legal writer	593,852	250,293
	Total generation 5		***3,809,892***	***2,081,646***
	Total		***87,126,198***	***17,307,907***

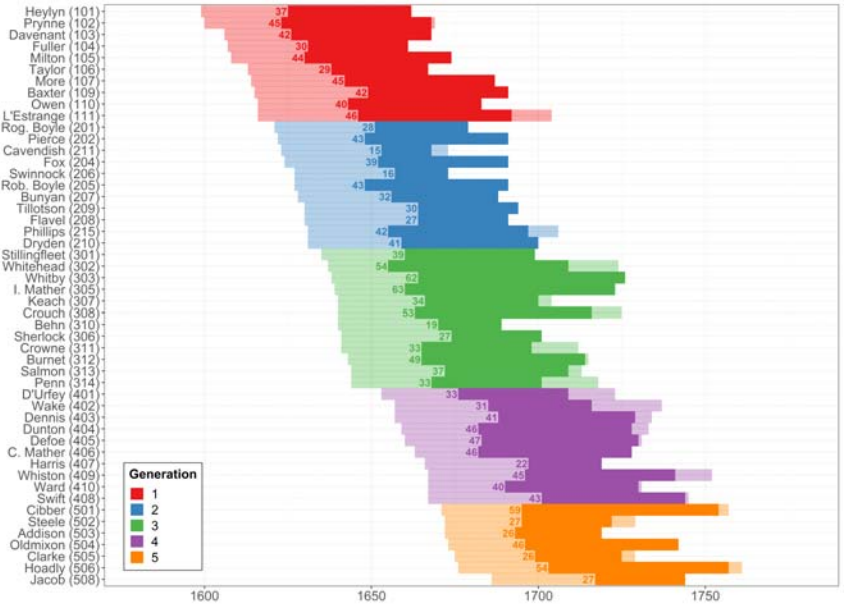


Figure 1: Lifespan and active career of EMMA authors; the darker areas represent the authors' active careers as covered in our corpus

Depending on the object one wishes to study, EMMA's unprecedented size may mean that a selection of the corpus will have to be used, as the case study might be too frequent to be examined exhaustively. One such case is presented in Anthonissen (ms.). For the purposes of that study, a principled selection of the available texts in EMMA was made, resulting in a 17-million-word sample (about 1/5 of EMMA's full size). Each author's active writing career was divided into five-year periods (starting from the earliest text) and for each period a sample of ca. 50,000 words was compiled, whenever possible consisting of a couple of texts across those five years rather than one large text.[2] Across periods, we aimed for a relatively constant genre distribution. This sample of EMMA, called EM for *EMMA Medium*, is fully compatible with EMMA (indices of the data points are co-referential) and can be used as a stand-alone corpus. Despite being more limited in scope, EM has the benefit of providing a more evenly distributed data set across individuals and periods (EMMA's goal, by contrast, was to include all available material per author and period). Because of this setup, EM lends itself well to the study of high- and mid-frequency phenomena. The EM corpus will become available with the second release of EMMA. The EM word counts are included in Table 1 for future reference.

3 Sources, formatting and markup

The texts in EMMA come from various sources (see references for details). The majority of texts was collected from the various clean text databases transcribed by the Text Creation Partnership (TCP), as well as from *Eighteenth Century Collections Online* (ECCO)[3]. The TCP text databases are EEBO-TCP (Phase I and Phase II)[4], ECCO-TCP, and Evans-TCP. Apart from ECCO-TCP, a couple of hundreds of texts were additionally retrieved from ECCO. Unlike the TCP selection, these texts were only available as uncorrected OCR. The noisy OCR was manually corrected with the correction tool provided by *18th Connect* (18thconnect.org). In addition to these major sources, a considerable number of texts were retrieved from various sources in the public domain. These include mainly Project Gutenberg (43 texts) as well as 30 more texts from twenty different web sources. The source of every text file is included in the general metadata sheet that is packaged with the corpus (cf. Section 4.1).

To ensure mutual compatibility, we converted the texts from various sources to a unified format. All texts encoded in Unicode UTF-8, and come in an XML format.[5] We maintained a minimal implementation of the Text Encoding Initiative (TEI5)[6], preserving all tags that might contain linguistically relevant information, such as text structuring tags (front, body, back) and highlight tags. Rich

metadata for the corpus are stored as XML-headers within the corpus files. Illegible characters and words were dealt with as follows: the *at*-sign stands for a missing word and underscores are used to mark illegible characters (e.g. *T_ing_* for *Things*).

All texts in the corpus have undergone two stages of automatic processing. The first stage involves tokenization, which was performed by means of an existing OpenNLP tokenizer (Apache Open NLP 2017) that we trained on the early modern section of the PENN-Parsed corpora (i.e. the PPCEME). This generally resulted in correct identification of tokens. The two major types of errors that occur relate to hyphenated compounds and apostrophes. Hyphenated compounds have sometimes been split up, when the training set did not contain compounds that were similar. In such cases, the hyphen ended up being treated as a separate token. A similar problem sometimes occurs with apostrophes. The tokenizer did for example not always correctly distinguish *'s* between a genitive marker (in which case it belongs to the preceding token) and the contracted form of *be* (in which case it is more likely considered a separate token). More generally, contracted forms of auxiliaries and the copula are more often merged with the preceding token than not (so <w>I'll</w> rather than <w>I</w> <w>'ll</w> – but both occur). These issues should be taken into account when querying the corpus.

The second automatic preprocessing stage deals with in-text language identification. On a macro-level, texts written in other languages than English have not been included in the corpus. On a micro-level, however, the corpus does contain some traces of code-switching. The most frequent foreign languages are Latin and French, and affected passages vary in size, ranging from (a part of) a sentence to entire paragraphs. In addition, the use of foreign languages varies considerably between the selected authors. This is troublesome, as it distorts the word count per author, affecting both relative and normalized frequencies. Although the accurate detection of foreign language spans in running text is not a trivial task, we have developed a language detection tool that tags the relevant passages in the corpus and as such provided us with estimates of the number of French and Latin words in each text. The tagging algorithm uses the log-likelihood probabilities of character trigrams in the various languages to create generative language models. In other words, for each sequence of three characters the algorithm has computed in advance an estimate of how likely such a sequence is to occur in either English, French, or Latin. At prediction time, it retrieves sufficiently long sequences (more than 10 words) that are significantly more likely to have been generated by the French or Latin language model than the English one and tags them accordingly. These estimates are included in the metadata

file, where they are subtracted from the raw word count. We opted for the ten-word threshold for pragmatic reasons. While tagging shorter passages did improve recall, it also yielded a lot of false positives (for instance words that exist both in English and French), which was detrimental to the algorithm's precision. As our main aim was to correct the word counts (as opposed to making the corpus suitable for code-switching research), we prioritized precision over recall and decided to invest in the accurate detection of longer sequences, at the risk of missing out on shorter ones.

Along with the corpus, we developed a custom interface that allows users to collaboratively query the corpus and annotate the hits. We released the interface as the open source package CosyCat (Collaborative Synchronized Corpus Analysis Toolkit) on GitHub (github.com/emanjavacas/cosycat), but the software is currently in alpha. CosyCat is naturally compatible with EMMA, but can also be used to query other corpora indexed by BlackLab[7] (de Does et al. 2017), or also, with minimal modification, corpora indexed by CWB/CQP (Evert and Hardie 2011). A fuller description of CosyCat is provided in Manjavacas and Petré (2017).

4 Metadata

4.1 Text metadata

Each text comes with a range of metadata. A metadata Excel sheet is packaged with the download of the corpus (for which see Section 7). This sheet provides information on EMMA corpus files and their respective source files. Columns A-T contain information concerning the corpus files, including text ID, author ID, title of the (main text), word counts, text date and genre classification. Columns V-Z list metadata retrieved from the source file (e.g. from EEBO and ECCO) and column U specifies whether the source file is open access.

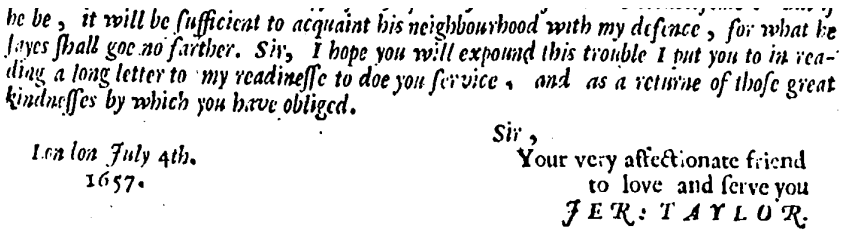
Apart from the word counts, text-specific metadata are also stored in the XML **<header>** element in EMMA's corpus files. Most of the information was automatically retrieved from the EEBO and ECCO databases and is retained under the **<sourceFile>** element. However, great care has been taken in verifying and complementing the metadata, especially date and authorship. We have also added a primary genre classification. We used XPaths to extract parts of texts that should either be retained or excluded in the author corpora, thus using text (rather than the printed volume) as the basic unit of our corpus. Metadata added by the *Mind-Bending Grammars* team are attached to the header under **<corpusFile>**.

The verification of metadata has been carried out along a number of dimensions. First, publication dates of the digital text's metadata were verified. This was done by looking for first prints in the entire EEBO database (including scans for which no transcription exists) as well as the *English Short Title Catalogue*[8] to see if an earlier edition was extant. For plays the date of the first performance was generally selected as corpus text date, as it was not uncommon for plays to be published only years later. Second, a first set of duplicates was automatically identified by means of *SpotSigs*, a robust algorithm for the identification of near-duplicates (Theobald et al. 2008). This automated procedure was complemented with a manual search for duplicates after compilation, on the basis of an inspection of the context around occurrences of a selection of mid-frequent patterns. This way it was also possible to identify partial duplicates, such as sermons that were printed both separately and as part of a collection. At the level of individual texts, several measures were taken to identify and extract parts with specific metalinguistic requirements, including the identification of non-authorial material. First, texts were split up into body, front and back, where front and back were generally ignored as they contain tables of contents, advertisements, indices, and the like, which we did not want to include as running text. Second, collections and volumes were scrutinized because they commonly lump together texts from various authors and various genres. We assigned XPaths to uniquely identify and extract those bits written by authors in our selection. The same procedure was used to assign specific genre labels.

By way of illustration of how all of this corrected information is represented in our corpus, Figure 2 shows the XML header of the work entitled "Certain letters of Henry Jeanes minister of Gods word …". In <sourceFile>, which refers to the original file in EEBO, the author is indicated as Henry Jeanes. However, one particular letter in this volume is written by one of our authors, Jeremy Taylor, as can be inferred from the signature in Figure 3. The letter also has a dateline (1657), which deviates from the publication date in the source file (1660). The <corpusFile> therefore lists 1657 as the correct date and specifies that the date was taken from a dateline. The letter was extracted by means of XPaths, so that of this volume only Jeremy Taylor's letter is retained in Taylor's corpus.

```
<?xml version="1.0" encoding="UTF-8"?><mbg>
  <header>
    <sourceFile>
      <title>Certaine letters of Henry Jeanes minister of Gods word at Chedzoy and Dr. Jeremy Taylor concerning a passage of his, in his further explication of originall sin.</title>
      <author>Jeanes, Henry</author>
      <publication country="United Kingdom" imprintLocation="Oxford" imprintPublisher="Printed by Hen. Hall for James Good, 1675." place="Oxford" pubDate="1660"/>
      <scan EEBOId="D00000119311310000" availability="Restricted" copyFrom="Union Theological Seminary (New York, N. Y.) Library" imageSet="51124" numPages="48" reelPosition="Wing / 816 :21" tcpId="A46697"/>
      <language>English</language>
      <biblInfo>Wing J504; ESTC R202621</biblInfo>
      <physicalDesc>[4], 48 p.</physicalDesc>
      <keywords>
        <keywords>Sin, Original.</keywords>
      </keywords>
      <notes>
        <note>Reproduction of original in Union Theological Seminary Library.</note>
      </notes>
    </sourceFile>
    <corpusFile>
      <sourceFile>/home/corpora/source/tcpii/utf/1/11931131.xml</sourceFile>
      <docId>11931131.0</docId>
      <author generation="1" id="106">Taylor, Jeremy</author>
      <date source="dateline">1657</date>
      <xpath>/EEBO/TEXT/BODY/DIV[1]/LETTER[2]</xpath>
      <genre>letters</genre>
      <PTC>NA</PTC>
      <textForm>NA</textForm>
    </corpusFile>
  </header>
```

Figure 2: XML header



be be, it will be ſufficient to acquaint his neighbourhood with my defence, for what he ſayes ſhall goe no farther. Sir, I hope you will expound this trouble I put you to in reading a long letter to my readineſſe to doe you ſervice, and as a returne of thoſe great kindneſſes by which you have obliged.

London July 4th. 1657.

Sir,
Your very affectionate friend
to love and ſerve you
JER: TAYLOR.

Figure 3: Metadata verification

As the dating of texts is crucial for the purpose of a corpus that allows lifespan research, efforts were also made to integrate information from various primary and secondary sources to ascertain a correct assessment of dating. For instance, Margaret Cavendish was first not selected for inclusion in the corpus, because of

dating issues relating to her two major published collections of plays. Initially it was not clear to us whether these collections brought together plays from various periods in her career, themselves undated (none of them were ever performed). This would make inclusion of these works problematic. However, based on circumstantial evidence kindly provided to us by James Fitzmaurice, we were satisfied that these collections had actually been created in a concentrated writing effort shortly before publication, as the following quote from Cavendish testifies:

> But my poor Playes, like to a common rout, Gathers in throngs, and heedlesly runs out, Like witless Fools, or like to Girls and Boyes, Goe out to shew new Clothes, or such like toyes: This shews my Playes have not such store of wit, Nor subtil plots, they were so quickly writ, So quickly writ, that I did almost cry For want of work, my time for to imploy. (Cavendish. 1662. *Playes written by the thrice noble, illustrious and excellent princess, the Lady Marchioness of Newcastle.*)

4.2 Genre classification

Another type of contextual enrichment is genre classification. Genre balance in itself was not a primary criterion, but the corpus contains considerable amounts of text from a wide range of genres that were common in the 17th century. The following are represented by at least 50,000 words in every generation: biography, conference, drama, hymns and psalms, legal texts, letters, footnotes, poetry, prayers, scientific texts, sermons, songs, and speeches. The current classification is inspired by the systems used in the ARCHER and Helsinki corpora, and has been double-checked by comparison with an automatic genre classification tool.[9] The classification is still preliminary, and at times remains underspecified. Further revision is planned for a future release. In what follows, we discuss the principles that underlie the current classification.

Genre classification was carried out on three levels (cf. Table 2). The most basic, formal level of distinction is **text form**, where we distinguish between prose and verse. Content-wise, a second level was assigned of **prototypical text categories** loosely inspired on a similar distinction used by the compilers of the Helsinki Corpus. We restricted this level to a broad three-way division into imaginative, non-imaginative and religious texts, grouping related genres in terms of topic and intended audience. Examples of the types of texts that were assigned to these categories can be found in Table 3. The third and most fine-grained classification is the label **'genre'**. Here we distinguished between the predominant written genres in the period under investigation (see Table 4). Our

goal was to provide EMMA users with a maximally informative label, which allows for an in-depth analysis of specific genres. This yields a rather substantial number of genre categories with many more subclassifications, yet we decided to keep the classification as fine-grained as possible since various genres may be easily lumped together in the data analysis stage if needed.

Table 2: Genre classification on three levels

CLASSIFICATION	XML HEADER NAME	PRACTICE
Text form	textForm	General distinction between: – prose – verse
Prototypical text category	PTC	General distinction between: – imaginative – non-imaginative – religious
Genre	genre	Labels are as specific as possible (but can of course be merged during data analysis if you are not interested in specific subsets of genres). Subcategories have an underscore.

Table 3: Prototypical text category

PROTOTYPICAL TEXT CATEGORY	[text types]
imaginative ~ fiction	Fiction, romance, drama, poetry, etc.
nonimaginative ~ non-fiction	Nonimaginative narratives and descriptive and/or argumentative texts on non-religious matters e.g. history, biography, memoirs, treatise, essay, document, law, handbook, science, philosophy, education, personal correspondence (non-religious), diary, travelogue, etc.
religious texts	Religious instruction, e.g. treatise, essay homily, rule, sermon, catechism etc. All other texts on religious matters: relation of church and state, episcopacy, religious persecution, religious aspects of secular matters (e.g. theatre, conduct of life, women, etc.), religious texts in verse (poems, hymns, prayers, etc.), religious letters, etc.
[combination of the above]	Overlap is allowed for: – if the text contains **various text types**, e.g. a letter and a poem – in the case of **biographies/histories** of religious persons/institutions, which are labelled non-imaginative+religious – if a text deals with **religious aspects of secular matters** such as theatre, life in the colonies, business etc.: non-imaginative+religious – if a text deals with an **historical event connected to religion** (e.g. Popish Plot) or **matters of church and state government** (rather than church alone): non-imaginative+religious
miscellany	if genre is classified as 'miscellany', prototypical text category unclear
undetermined	prototypical text category could not be determined

Table 4: Genre

GENRE [label]	[description]	[subcategories]
science	scientific texts	science science_chemistry science_geography science_mathematics science_medicine science_physics
legal	legal texts	legal
letters	letters	letters letters_monitory letters_pastoral (= pastoral letters and charges)
sermons	orations or lectures by a member of the clergy	sermons_election sermons_execution sermons_fast-day sermons_funeral
satire	satires	satire
fiction	imaginative narrative prose	fiction
drama		drama drama_comedy drama_farce drama_masque drama_opera drama_prologue/epilogue drama_tragedy drama_tragicomedy
poetry	poetical work	poetry poetry_burlesque poetry_elegy poetry_epic poetry_epigram poetry_heroic poetry_miscellany (various types of poems) poetry_occasional (panegyric poems, congratulatory poems, funeral poems, etc.)
songs	songs, ballads	songs
hymns/ psalms	religious songs	hymns/psalms
catechism	religious instruction in question-answer form	catechism

biography/ memoirs	biographies, memoirs, memories and accounts of the life and death of a particular person	biography/memoirs
dialogue/ conference	dialogues, conferences, interview and discussions with turn-taking	dialogue/conference
speech	speeches or talks	speech
prose	generic label for argumentative and/or descriptive prose (they may or may not belong to the text category 'prose', which serves a different purpose of differentiating with verse.)	large prose subcategories (sermons, legal, scientific texts) have a separate genre label
v	various other minority genre categories	v_advertisement v_fable v_parable v_testimony v_prayer
miscellany	mix of different text types	miscellany
undetermined	genre is unclear	undetermined

4.3 Author metadata

In addition, a more extensive author metadata database is underway with rich biographical information on each author. This includes information on birth and death dates, birth place, social circles, political and religious orientation. It also includes quantifiable social network information and the mobility history of each author. The metadata database is currently in alpha. The database will be made publicly available together with a first revision of the corpus itself at the end of the *Mind-Bending Grammars* project in 2020.

Figures 4 and 5 illustrate how this metadata can be sensibly quantified. Both figures show the social network connections between the individuals in EMMA. The first visualizes live connections between the individuals, trying to provide an approximate answer to the question: how often did they meet each other in real life? The second visualizes the citation network of our individuals: how often did they cite each other? Of course these networks should not be considered as self-contained autonomous wholes. Rather they represent snapshots of larger networks (such as the literary scene, the community of clergymen, London), and should be interpreted in this light. Similar to the approach found in Bergs (2005: 55–70) and Sairio (2009), we have assigned weights to network ties, but our approach differs in assigning tie strength in a more data-driven way.

The different procedure is motivated by the different nature of the sources. Both Bergs (2005) and Sairio (2009) analyse correspondence, where influence between informants is tested on the basis of letters they wrote to each other, and is mostly concerned with interactional accommodation. The connections that we can establish between informants in EMMA are generally of a more indirect nature. In integrating actual mentions rather than establishing tie strength on the basis of a global biographical and social profile, the aim is to inform analyses of converging (or diverging) behaviour between individuals that resulted from interactions that are essentially invisible to us. Examples are the adoption of someone's idiosyncrasies by a friend (live) or an admirer (in reading), or shared language use typical of one of the communities represented in EMMA, such as that of the royal chaplains.

Both the networks visualized in Figures 4 and 5 were calculated using the same methodology. For each individual we[10] counted the number of times any of the other individuals was mentioned in (i) their biography pages on Wikipedia and the *Oxford National Dictionary of Biography*[11] combined; (ii) their own written work in EMMA. For each mention it was decided what the kind of connection was. This decision was always informed by the context. For instance, from the following reference of Gilbert Burnet (312) to Richard Baxter (109), we can infer that they knew each other only from certain society meetings, but not at all closely. Hence Burnet's surprise that Baxter has witnessed against him in a fairly serious allegation of treason.

> The Witnesses cited against me are first [...] and for the last, Mr. Baxter, I have had no Correspondence at all with him these two and Twenty Years; unless it was that once or twice I have met him by accident in [Visit] in a third place, and that once about six Years ago I went to discourse with him concerning a matter of History in which we differ'd; but as all our Conversation at that time was in the presence of some Witnesses so it was not at all relating to matters of State. (Burnet. 1687. *Six papers by Gilbert Burnet.*)

Consequently, this mention is tagged as 'society'. Connection types were based on the kind of connections that were attested in our sources, and include (for example) (paper) friend, (paper) ally, (paper) opponent, classmate, colleague, professional collaboration, professional connection, family, Quaker, supporter/supportee, admirer/admired, influencer/influenced, imitator, audience (context of preaching), reader, or reviser. We then ranked these types by assigning weights. This procedure was motivated by the likelihood of a misbalance between the frequency of mentions in our sources and the frequency of contact

in reality. Family ties and friendships will generally be less reported on in the sources we have than, for instance, opponents, allies, or professional collaborators. To compensate for this, family and friend mentions received a weight of 2. Similarly, citations of admiration received a slightly higher weight of 1.25 than neutral citations (weight of 1), under the assumption that admiration triggers imitation. Indirect or distant mentions (e.g., someone repeating some rumour about someone else) received a weight of 0.75. These weights are currently assigned intuitively, but generally in line with the more sociologically informed study by Sairio (2009).

After assigning these weights to each mention, the weighted numbers were then added up (for each of the categories 'live' and 'citation/paper mention'). The resulting number was normalized by dividing it by that individuals' corpus size and size of the biography. Finally, normalized numbers were divided into ten ranked bins. These ranks for each (directed) pair of individuals were then, finally, fed into the Force Atlas 2 algorithm available in the Gephi software package (Bastian et al. 2009).

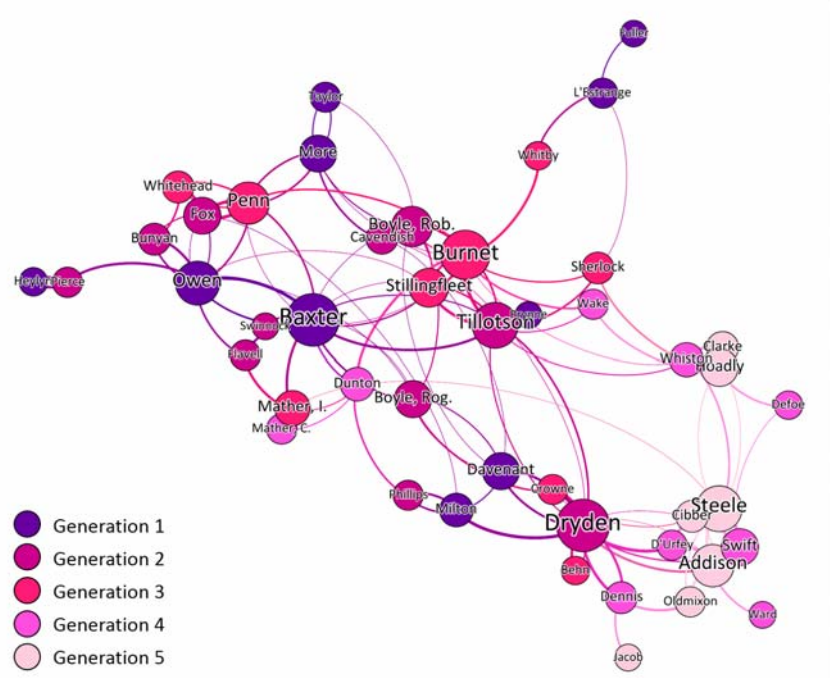


Figure 4: Network of live social connections between EMMA informants

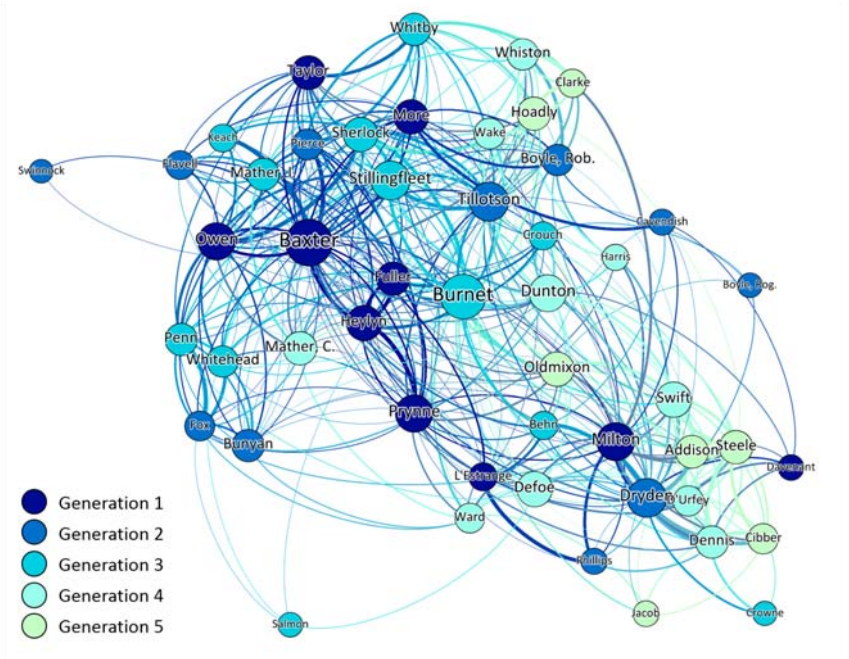


Figure 5: Network of citation and mention connections between EMMA informants

The most obvious difference between the two networks is their density. This is partly the natural consequence of the fact that early and late generations often cannot have met, because they were never contemporaries (at least not as adults). Partly it may be the result of the lack of appropriate sources. While a citation network based on a near-exhaustive sample of published work can be expected to be fairly representative, live connections are harder to accurately cover with these resources. It would be interesting to add the information of their private unpublished correspondence, and in general a more principled methodology can be envisaged, but designing and implementing such a methodology would require a separate research project beyond our current scope.

Apart from the difference in density, there are also obvious parallels. In both networks the central figures are among the most influential leaders in their circles. Richard Baxter, John Tillotson, Edward Stillingfleet and Gilbert Burnet were the most important religious leaders in their time. John Dryden does credit to his reputation (the 17th century is also commonly referred to as the age of

Dryden). Other influential authors such as John Milton or Addison and Steele are also central, but less so, perhaps because they are born in the first and last but one generations, whereas Dryden is situated in the middle generation. Finally, while the citation network shows that all individuals in EMMA are interrelated on paper, the live network does not contain all our authors. Some authors, such as William Salmon (313), a doctor derided by colleagues as "the King of Quaks" (Hanson 2009: 118), were apparently quite isolated from the social circles to which most of EMMA's authors belonged.

5 Research opportunities and examples

In this section we will briefly show the kind of research questions that can be tackled with EMMA. Sections 5.1 and 5.2 present two new small case studies. Section 5.3 provides a summary of research that has already been carried out with the help of EMMA.

5.1 Case study 1: Strong verbs

The first case study that we will use to illustrate the potential of EMMA is that of ablaut variation in strong verbs. We have analysed the ablaut choices in the past tense and participle (*a* or *u*) as made by nine of the authors[12] (from three generations) in EMMA for the verbs *begin, cling, drink, fling, ring, sing, shrink, sink, sling, slink, spin, spring, stink, sting, string, swim, swing, wring*. While a fair few of these no longer show variation in Present-Day English (Anderwald 2011), in Early Modern English more of them still did,[13] so we included this wider range. Anderwald (2011) shows that from the 18th century onwards prescriptive grammarians put much effort in promoting for many of these verbs a three-way distinction between present, past and participle, as in *drink – drank – drunk*. These efforts were essentially based on a (mistaken) Latin ideal, as Latin as a rule has distinct forms in past tense and participle. Anderwald (2011: 91) furthermore points out that "these verbs span frequency bands from the medium-frequent (*drink*, *begin*) to the quite rare (*slink*, *spring*)", and that the limited size of corpora makes it impossible to establish whether "standardization had already set in for these verb forms". She illustrates this for the past tense form *drunk* (rather than the more frequent *drank*), which only occurs once in the Helsinki Corpus outside Old English, and once in ARCHER 1, within the period 1700–1750 (Biber, Finegan and Atkinson 1994). Therefore Anderwald starts her own analysis of the relation between prescriptivism and linguistic reality with an analysis of prescriptive grammars from the 19th century, comparing this with current usage. She concludes that initially there is a strong attempt to force the

three-way system onto most of the verbs at issue, but gradually prescriptive grammars show a little more tolerance towards *u*-forms in the past tense, arguably reflecting a persistent linguistic reality.

EMMA provides an excellent resource to dig deeper into some of the issues involved in the variation in strong verb inflections. First, the corpus size largely makes up for the medium or rare-frequency ranges of these verbs. Instead of the one example of past tense *drunk* in the Helsinki Corpus, the nine authors that we examined (constituting less than a fifth of EMMA) already yield twenty-seven instances of past tense *drunk*. To get robust results on an individual basis we still decided to aggregate most verbs, commenting on individual verbs where required. The exception is *begin,* which is far more frequent than the others, and most distinct, as all authors prefer a three-way ablaut distinction for this verb. One source of concern, which is less relevant when studying syntactic change, is the potential influence of typesetters, correctors, or publishers on spelling standardization (Howard-Hill 2006). While we do not deny this impact, ablaut is not generally mentioned among the categories that are being corrected in the 17th century. Early modern errata lists also do not seem to contain ablaut corrections, even though concerns about a similar phenomenon such as number agreement led to plural *was* occasionally being changed into the 'correct' plural *were* (Lepo 2018: 60). The rather liberal, and therefore presumably faithful attitudes towards ablaut seem to be confirmed by the considerable amount of ablaut variation within single documents in EMMA. Variation is also consistently present across printers. For Benjamin Keach, for instance, we identified fifteen different printers, but there is no evidence that any of them skewed Keach's data in a specific direction.

A second advantage of EMMA is that the different backgrounds of the individuals allow us to investigate a bit closer the idea that the three-way-distinction is inspired by Latin grammar. While our period predates the surge of prescriptive grammars in the second half of the 18th century, the authors of these grammars arguably share some essential background with our authors. Specifically, if these authors are really inspired by Latin, this means that they will have spent a great deal of time studying Latin, and considering the properties of their own language as well. While Latin was a subject of study in grammar schools, a more reflective study of the language as well as of English is likely to have occurred primarily at university. In renaissance Europe, a university degree largely boiled down to a thorough study of the classical languages and cultures. It is therefore not unreasonable to assume that early modern authors who took a university degree had much more opportunity to discern possible underlying parallelisms between English and Latin than other authors. The role of higher

education in the linguistic behaviour of the informants in EMMA is also observed in another study by Standing and Petré (*subm.*). In that study it is shown that only authors without a university degree show significant lifespan change in how they use clefts. This seems to suggest that authors with a university degree had already spent time on crystallizing their grammatical behaviour, not unlikely in the context of their university studies. In the current case of ablaut variation, the input of Latin may be expected to be even more specific. If it turns out that authors with a university degree make use of a three-way distinction more often than those without a degree, this would be an indication of how the prescriptive tradition emerged out of their language customs.

Figures 6 and 7 visualize the distribution of *a* and *u*-ablaut across preterite (vbd[14]) and past participle (vbn) in our authors, grouped by generation. Figure 6 represents individuals with a university degree. Figure 6a (first row) shows the aggregate results for all verbs except *begin*. Figure 6b (second row) zooms in on *begin*. The Kendall tau-b rank correlation test was used to establish how much discordance there is between the generations, so as to assess any trends. As it turns out, informants with a university degree show a significant increase across the three generations in the use of the *a*-ablaut in the past tense at the cost of *u*-ablaut, towards a more categorical three-way system (type *sing – sang – sung* rather than *sing – sung – sung*). Judging by Figure 6b it appears that *begin* shows a categorical three-way distinction throughout. *Begin*, then, may have been a kind of model that was increasingly extended to other verbs. Figure 7 shows the ablaut distribution in authors without a university degree. Unlike the highly educated group, this second group does not show a steady increase of *a*-ablaut. Their use of *a*-ablaut is also consistently lower than that of the highly educated group. If anything, this group tends in the opposite direction, towards increased use of *u*. Indeed, while not significant for the aggregate group, there is a significant increase of *u*-ablaut in the past tense of *begin*.

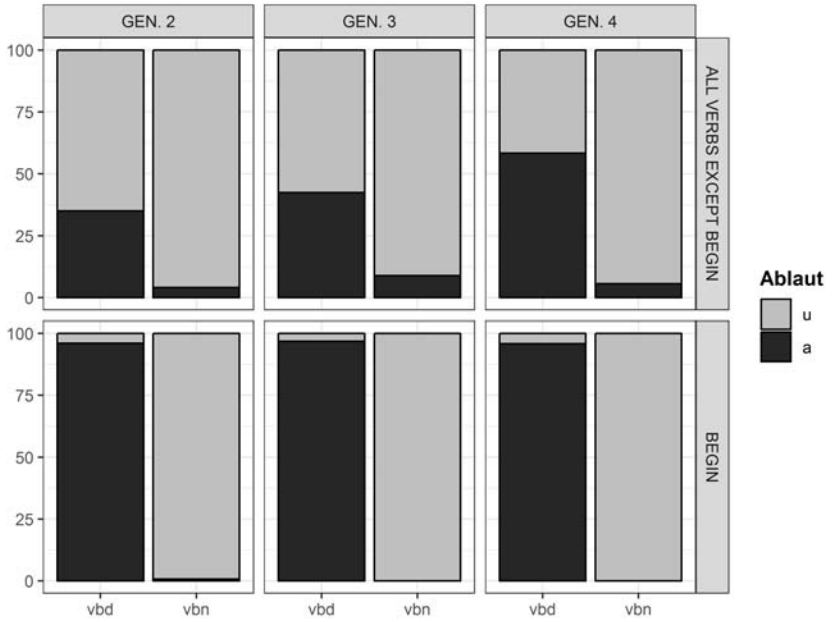


Figure 6: Authors with a degree, by generation; (6a) all verbs except begin (tau-b=0.16; p=0.004; n=318); (6b) begin (tau-b=0.002; p=0.9; n=902)

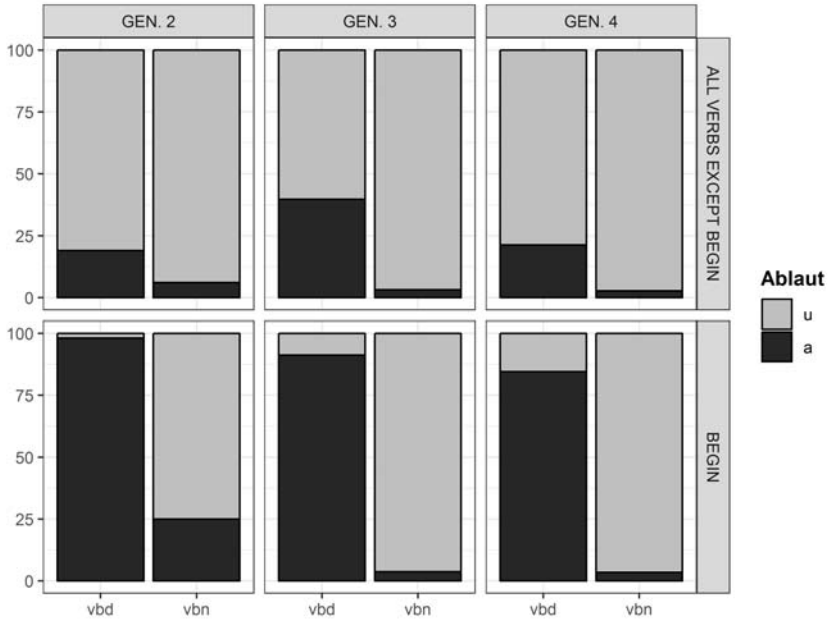


Figure 7: Authors without degree, by generation; (7a) all verbs except begin (tau-b=0.029; p=0.57; n=370); (7b) begin (tau-b=0.14; p<0.001; n=571)

The apparent extension of the three-way distinction in the highly educated group rarely if ever occurred in a categorical fashion. *Drink* is a good example. Unlike most other verbs, where *u*-ablaut tends to dominate across the board, *drink* generally preferred *a* in both the past tense form and the participle. John Dryden and Jonathan Swift, authors known for their linguistic fastidiousness, reveal attempts at a three-way distinction but without complete success. Dryden uses 79 percent *a* in the past tense and 71 percent *u* in the participle (n=38). Swift, two generations later, appears to be near-categorical, with seven times *a* and no *u* in the past tense and two *u*-s against one *a* in the participle, with however only three instances in total. Two other highly educated authors, William Sherlock and William Wake, show a consistent two-way distinction, using *a*-ablaut for both past tense and participle. One cannot therefore really claim that the highly-educated had a clear perception of the way they believed things should be. Nevertheless, their familiarity with languages such as Latin may have influenced

their use of English, which displayed probabilistic influences that gradually crystallized into categorical distinctions. At the same time the informants without a degree continued to prefer *u*-ablaut in the past tense, and even show signs, in the verb *begin*, of extending *u*-ablaut across the board. It seems plausible that it was precisely this growing discrepancy between these two groups – a discrepancy which may have been even larger with non-writers – that turned this category of strong verb inflection into an index (in the sense of Eckert 2008) of learning (and any social associations that go with it), and as such fed an ever stronger prescriptive reflex in the 18th century.

Zooming in on lifespan change, among all nine authors only Benjamin Keach shows significant lifespan change. Figure 8 shows his ablaut choice in the past tense in the first and second halves of his career. While starting out with a low usage of *a*, Keach catches up on this later in life (chi-square p-value=0.07). Such age-grading may in this case point to a continued effort by Keach to adapt to the language use of the educated elite. He even seems to overreach and fall victim to hypercorrection, as when he uses *sank* (twice) for instance, where most highly educated authors at that time generally have a two-way conjugation *sink – sunk – sunk*. Maybe Keach was motivated to increase his credibility as a serious author. Regardless, the observation that significant lifespan change only occurs with a writer who did not attend university further corroborates the findings in Standing and Petré (*subm.*).

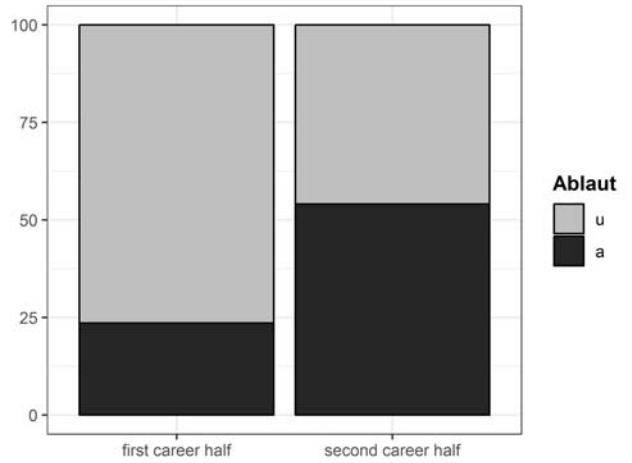


Figure 8: Ablaut choice in the past tense in Keach's early and late work

5.2 Case study 2: -s/-th in the language of Margaret Cavendish

The second case study focuses on a single individual only and concerns the morphological variation in the third person singular between *-s* and *-th*. We selected this case study because it nicely illustrates a second important social factor besides the role played by education illustrated above. This factor is more demographic in nature and relates to the density of the network of native speakers in an individual's environment.

It has been shown repeatedly that members of larger communities (and therefore the larger communities themselves) evolve faster linguistically (Milroy and Milroy 1997; Nevalainen 2000; Trudgill 2011). Nevalainen (2000) shows that specifically for English, London is at the forefront of most early modern innovations. Conversely, small communities with strong bonds tend to stick to their own local lects. Cavendish's use of *-s* and *-th* in the third person singular suggests that even a single language user may change radically when they shift communities. To examine this we analysed the choice of inflection in her work across time for the verbs *advise, cause, close, concern, consist, feel, keep, laugh, pass, produce, promise, rise, seize, talk*. These verbs were selected to cover a good range of mid-frequency verbs of which it may be assumed that they were shifting to *-s* at more or less similar rates. An equal amount of verbs ending and not ending in a sibilant is included, as this phonotactic property may have played a role in adoption rates of *-s* as well (see e.g. Walker 2017). Verbs such as *do* or *have*, which are known to stick to *-th* for an unusually long time (Nevalainen, Raumolin-Brunberg and Mannila 2011: 11), have been deliberately excluded.

Figure 9 shows the results. While in most of the years for which we have data Cavendish shows a clear preference for the incoming form in *–s*, the years 1655 and 1656 (and to a lesser extent 1662) stand out. In these years her published work shows a marked preference for *-th*. Why this marked difference? We believe her behaviour can be explained by looking into her biography more closely. Margaret Cavendish, née Lucas, married William Cavendish, Marquis (later Duke) of Newcastle in 1645 (Fitzmaurice 2004). There are two important facts about their marriage. First, the Duke of Newcastle was 31 years older than Margaret, and therefore belonged to a generation where the use of *-th* still prevailed. Second, they spent a good deal of their lives together in exile in France (Paris) and the Netherlands (Antwerp, Rotterdam). As a royalist, the Duke of Newcastle decided to leave England when Cromwell seized power during the Civil War. It was in Paris that he met Margaret and that they married. They then spent most years until 1660 – when the king was restored to the throne – in exile. While they will have had contact with other exiles, their primary source of

exposure to English must have been each other. It is remarkable, then, that Margaret's use of *-th* peaks precisely in this period of isolation, and it seems likely that she accommodated her language to that of her husband. While her husband does not have a very large corpus in EEBO, it is indeed the case that in his publications he prefers *-th* throughout. The seemingly exceptional year 1653 further corroborates this hypothesis. Between the winter of 1651 and the summer of 1653 Margaret spent about eighteen months without her husband back in London. It is during this time she wrote and published the *Philosophicall fancies and* also *Poems, and fancies,* in both of which *-s* is predominant.

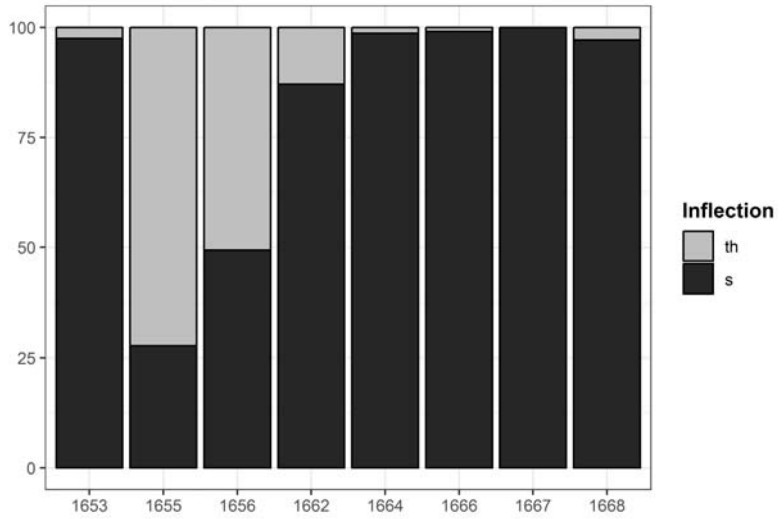


Figure 9: Cavendish's choice of inflection across the lifespan (n=848)

In sum, EMMA also provides the means to test how embeddedness in a community and shifts in such embeddedness impact on language behaviour. The relation between innovative behaviour and the urban environment of London has also been shown in a larger scale study based on EMMA by Van de Velde and Petré (2017), on the grammaticalization of *be going to* INF.

5.3 Highlights of EMMA-based research so far

The above sections provided two small illustrations of how EMMA can be used to assess the impact of mainly social factors on language change taking fully

into account the individual dimension. Besides these, various studies have made use of EMMA, either focusing on the interaction of cognitive and social factors, or zooming in further on the cognitive dimension. Within the first category, Van de Velde and Petré (2017) have shown how the grammaticalization of *be going to* INF is not only led by the London community, but individual users' position on the grammaticalization cline for this construction also turned out to strongly correlate with the age at which they came to live in London (if ever). Standing and Petré (*subm.*) is a study of how the cleft construction is undergoing functional and structural shifts in Early Modern English, as reflected in EMMA authors. In addition to intergenerational incrementation (i.e. the advancement of change between generations, Labov 2007: 346) it was found that certain individuals without a university degree change their use of clefts significantly over the lifespan, going from more diffuse usage to a usage that is more focused on a preferred function.

The interaction between cognitive plasticity and social embeddedness is worked out in Petré and Van de Velde (2018), a study on the early grammaticalization of *be going to* INF as reflected in EMMA. Evidence is accumulated that there is a difference between individuals who grew up before, during, and after the conventionalization of grammaticalized *be going to* INF. The first group adopts the novel construction only to a limited extent. The second (consisting of generations 2 and 3) has it from the start, but does not use it to the fullest initially, perhaps inhibited by social accommodation to the previous group. Some individuals within this group show significant growth in degree of grammaticalization over the lifespan. The third group has reached a ceiling and shows a maximal presence of the grammatical features of the construction at this stage. As a result they do not show any progress over the lifespan. Also taking *be going to* as a case study, Anthonissen and Petré (2019) have presented complementary evidence that (highly educated) monolingual speakers continue to participate in grammatical innovations across the lifespan (up to past the age of 60). People who adopt the innovation at a later age even show signs of reanalysing (rather than superficial pattern borrowing) their spatial schema of *be going to* INF into a more grammatical prospective future, in which *be going* is now analysed as an auxiliary. However, such reanalysis is increasingly constrained by pre-existing language habits in combination with the functional and/or formal distance between source and target construction.

Other studies concentrate on cognitive factors in constructional usage, such as their mutual association and changes within those associations within individuals. Standing, Strik and Petré (*subm.*) argue that intransitive and transitive uses of *get* with a complement (*He got himself/me ready* and *he got ready*) are inter-

related in individuals' minds. They also provide statistical support for a similar association between clefts and extraposition. In the first generation of authors their frequencies correlate significantly. However, this correlation is markedly weaker in the second generation, suggesting that clefts are coming into their own, emancipating from the weight management function they shared with extraposition. Related work on functional shifts in syntactic constructions is found in Anthonissen (2019), who shows how the two main functions of the NCI construction (*he was said to be a thief*) shift in relative frequency in individuals across their lifespan. Specifically those individuals who show lifespan change consistently shift towards higher use of the evidential function of the NCI, in line with what is going on at the communal level. It is important to note that these correlations and changes in them are not merely an aggregate phenomenon, but are shown in these studies to be recurrent in individuals from the same generation.

Finally, another set of studies concentrates on the link (or discrepancy) between individual and communal behaviour. Fonteyn and Nini (ms.), for instance, make use of conditional inference trees to identify hierarchies in the factors that determine the choice between nominal and verbal gerunds. They show that behavioural clusters are found among traditionally described factors (such as the gerund's function as direct object or prepositional phrase), but that these are overruled by a distinction between two types of individuals that cannot be further interpreted in terms of the traditional syntactic-functional literature. They also observe and warn against the way in which the distribution at the aggregate level may in fact be obliterating and in contradiction with what happens at the individual levels. Similarly, in a study on the rise of the prepositional passive, Anthonissen (ms.) demonstrates that regularities and trends that arise at the aggregate level of language (e.g. a steady increase in normalized frequency) conceal the complexity and unpredictability found at the individual level. Intermediate levels of abstraction, whereby for instance age cohorts or group membership are taken into account, may also reveal systematicity that is not apparent in individual behaviour. Anthonissen (ms.) furthermore shows that the minority group that exhibits a lifespan increase in line with the communal trend is also the group of authors leading the change.

What most of these studies share is that EMMA enables them to quantify linguistic phenomena or features that are below the frequency range that can be studied on the basis of letters, which has been a major source for lifespan research so far (e.g., Raumolin-Brunberg 2009), and do so for more informants than with smaller-scale corpora (e.g., Fitzmaurice 2004). Some studies, such as Fonteyn and Nini (ms.), examine high-frequency phenomena like the gerund,

but benefit from EMMA's size to carry out multi-variate analyses on an individual basis to classify individuals in types. Similarly, the higher frequencies of features that signal ongoing renegotiation of form-meaning relations, as is the case with most syntactic change, provide a more solid basis to gain insight in the cognitive mechanisms underlying changing language behaviour (as in Anthonissen and Petré 2019).

6 Future improvements

EMMA is currently available in its first release. Improvements and changes to the corpus will be accumulated in the future into occasional new releases, with their own version number. Care will be taken that the token indices of future versions are compatible with those of previous versions, so that it should be possible to largely automatically update any research databases based on EMMA. Future improvements are planned along two dimensions.

First, while we have put much effort and time in the accuracy of the metadata of the EMMA-corpus, given the complexity of the undertaking, it is only natural that improvements can still be made here. One aspect we intend to further improve and complete is the genre classification. The category 'prose' for instance, is currently rather broad, and as such not particularly useful for discrimination. Apart from genre, corrections and additions to dating and authorship are being made whenever we come across the right information. In between versions, these updates will be occasionally added to EMMA's download page. When a new version comes out, they will be integrated in that version.

Second, we have concrete plans for the near future to implement spelling normalization on EMMA. Normalized spelling will be made available as XML-attributes to the token, which will retain its original spelling. Normalization will be carried out with the University of Lancaster's VARD-tool,[15] but the tool has been tuned to better suit our particular corpus data. Spelling normalization is especially useful for automated (NLP) applications, but traditional corpus linguistics will benefit from the normalization too, as it diminishes the need to adjust the queries to accommodate for a multitude of spelling variants. In a similar vein, the normalized corpus will also be POS-tagged by the Early Modern English POS-tagger of the MorphAdorner package (Burns 2013).

7 Availability and contact information

A copy of the corpus can be requested at https://www.uantwerpen.be/en/projects/mind-bending-grammars/emma-corpus/. After registration a download

link will be provided. The majority of texts is currently already in the public domain. However, a minority (those from the source database EEBO Phase II) will only enter the public domain in 2020. During this transition, researchers from institutions with a subscription to EEBO-TCP Phase II can already download the complete EMMA corpus; an open access version without EEBO Phase II is available for those without subscription. The remainder of EMMA will be made available to all once the source texts have entered the public domain (around 2021).

Table 5 provides contact details and general facts about EMMA.

Table 5: EMMA fact sheet

Project leader	Peter Petré
Compilers	Peter Petré, Odile A. O. Strik, Lynn Anthonissen, Sara Budts, Enrique Manjavacas, William Standing, Emma-Louise Silva
Volunteers	Maria De Graef, Lutgarde De Haeck (main contributors), Diane Koek, BA and MA students from the University of Antwerp
Time of compilation	2015–2018
Size	90 million words (inclusive non-English text); 88.5 million (English only)
Language	English
Number of texts/samples	13,750
Period	1623–1757
Released	2018 (version 1.0)
Funding	H2020 - European Research Council (ERC) (Project ID 639008)
Corpus home page	www.uantwerpen.be/en/projects/mind-bending-grammars/emma-corpus/
Contact	emma@uantwerpen.be (corpus inquiries); peter.petre@uantwerpen.be (other)
CosyCat	github.com/emanjavacas/cosycat
Project website	www.uantwerpen.be/en/projects/mind-bending-grammars

We would be grateful for users to cite this article when using EMMA. The corpus itself can additionally be cited as follows:

Petré, Peter; Lynn Anthonissen; Sara Budts; Enrique Manjavacas; Emma-Louise Silva; William Standing; and Odile A.O. Strik. 2018. *Early Modern Multiloquent Authors (EMMA), release 1.0*. University of Antwerp, Linguistics Department. Online: https://www.uantwerpen.be/en/projects/mind-bending-grammars/emma-corpus/.

Acknowledgments

The research reported on in this paper is part of the project *Mind-Bending Grammars*, which is funded by the ERC Horizon 2020 programme (Project ID 639008; www.uantwerpen.be/mind-bending-grammars/), and is hosted at the University of Antwerp. Both institutions are hereby gratefully acknowledged. We would also like to thank two anonymous reviewers for their generous comments and suggestions. Finally, special thanks go to our two most relentless volunteers Maria De Graef and Lutgarde De Haeck, who between them corrected almost 6,000 pages of OCR-ed text on a voluntary basis.

Notes

1. Due to a lack of more suitable candidates, this includes John Milton in generation 1. Milton wrote some drama and masques, although this is a relatively small part of his oeuvre and is often also not very comparable to that of most of his contemporaries.
2. Deviations were allowed for when the material was not available in those periods or if the author wrote relatively few, but large works. In the latter case, we generally included more words per period by selecting parts of the larger work to make up for distributional gaps.
3. Eighteenth Century Collections Online (quod.lib.umich.edu/e/ecco)
4. Early English Books Online (eebo.chadwyck.com)
5. A plain text format may be generated upon request in the future.
6. See www.tei-c.org/guidelines/p5.
7. github.com/INL/BlackLab
8. http://estc.bl.uk
9. Developed in collaboration with MA student Arthur Nieuwland during an AI internship at *Mind-Bending Grammars*.
10. The core team and MA student Géraldine Vandamme who did most of the first draft analysis of the EMMA citations.
11. www.oxforddnb.com
12. These are: George Fox (204), John Flavell (208), John Dryden (210), William Sherlock (306), Benjamin Keach (307), Aphra Behn (310), Thomas D'Urfey (401), William Wake (402), and Jonathan Swift (408).
13. As is explained in some detail in Anderwald (2011), this variation finds its origin in the Germanic ablaut system, which distinguished between the past tense singular on the one hand and the past tense plural and participle on the other. This system had already more or less broken down before the start of our data, and our authors' variation may be seen as reflecting ongoing

attempts to realign the ablaut distinction with the past versus participle distinction.

14. These are manually annotated POS-tags, for maximal accuracy. Automatic POS-tagging has not yet been implemented (cf. Section 6).
15. ucrel.lancs.ac.uk/vard/about/

References

Anderwald, Lieselotte. 2011. Norm vs. variation in British English irregular verbs: The case of past tense *sang* vs. *sung*. *English Language and Linguistics* 15: 85–112.

Anthonissen, Lynn and Peter Petré. 2019 (forthcoming). Grammaticalization and the linguistic individual: new avenues in lifespan research. To appear in *Linguistics Vanguard* (Special Issue: *Language and Aging*).

Anthonissen, Lynn. (Manuscript). Cognition in construction grammar. *Cognitive Linguistics* (Special issue: *Constructionist Approaches to Individual Grammars*).

Anthonissen, Lynn. 2019 (forthcoming). Constructional change across the lifespan: The nominative and infinitive in early modern writers. To appear in K. Bech and R. Möhlig-Falke (eds.). *Grammar – discourse – context: Grammar and usage in language variation and change* (Discourse Patterns). Berlin: De Gruyter Mouton.

Apache OpenNLP. 2017. The Apache Software Foundation. https://opennlp.apache.org

Archer, Ian W. 2000. Social networks in Restoration London: The evidence of Samuel Pepys's diary. In A. Shepard, P. J. Withington and P. Withington (eds.). *Communities in early modern England: networks, place, rhetoric*, 76–94. Manchester: Manchester University Press.

Bastian, Mathieu, Sebastien Heymann and Mathieu Jacomy. 2009. Gephi: An open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*. www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

Beckner, Clay, Richard Blythe, Joan Bybee, Morten H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman and Tom Schoenemann. 2009. Language is a complex adaptive system. *Language Learning* 59: 126.

Bergs, Alexander. 2005. *Social networks and historical sociolinguistics: Studies in morphosyntactic variation in the Paston Letters (1421–1503)* (Topics in English Linguistics 51). Berlin: Mouton de Gruyter.

Biber, Douglas, Edward Finegan and David Atkinson. 1994. ARCHER and its challenges: Compiling and exploring A Representative Corpus of Historical English Registers. In U. Fries, G. Tottie and P. Schneider (eds.). *Creating and using English language corpora*, 1–14. Amsterdam: Rodopi.

Burns, Philip R. 2013. *MorphAdorner v2: A Java library for the morphological adornment of English language texts.* Evanston: Northwestern University. https://morphadorner.northwestern.edu/morphadorner/download/morphadorner.pdf.

Bybee, Joan L. 2010. *Language, usage and cognition.* Cambridge: Cambridge University Press.

Dąbrowska, Ewa and James Street. 2006. Individual differences in language attainment: Comprehension of passive sentences by native and non-native English speakers. *Language Sciences* 28: 604–615.

Dąbrowska, Ewa. 2015. Individual differences in grammatical knowledge. In E. Dąbrowska and D. Divjak (eds.). *Handbook of cognitive linguistics*, 649–667. Berlin: De Gruyter Mouton.

de Does, Jess, Jan Niestadt and Katrien Depuydt. 2017. Creating research environments with blackLab. In J. Odijk and A. van Hessen (eds.). *CLARIN in the Low Countries*, 245–257. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbi.20. License: CC-BY 4.0

ECCO = *Eighteenth Century Collections Online*. quod.lib.umich.edu/e/ecco.

ECCO-TCP = *Eighteenth Century Collections Online – Text Creation Partnership.* www.textcreationpartnership.org/tcp-ecco.

Eckert, Penelope. 2000. *Linguistic variation as social practice*. Oxford: Blackwell.

Eckert, Penelope. 2008. Variation and the indexical field. *Journal of Sociolinguistics* 12 (4): 453–476.

EEBO = *Early English Books Online*. eebo.chadwyck.com.

EEBO-TCP = *Early English Books Online – Text Creation Partnership*. www.textcreationpartnership.org/tcp-eebo.

Ellis, Nick C. 2011. The emergence of language as a complex adaptive system. In J. Simpson (ed.). *The Routledge handbook of applied linguistics*, 654–667. New York: Routledge.

Evans-TCP = *Evans Early American Imprints – Text Creation Partnership.* www.textcreationpartnership.org/tcp-evans.

Evert, Stefan and Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics Conference 2011, Birmingham, 20–22 July.* Paper #153. https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-153.pdf.

Fitzmaurice, James. 2004. Cavendish [née Lucas], Margaret, duchess of Newcastle upon Tyne (1623?–1673), writer. *Oxford dictionary of national biography.* Oxford: Oxford University Press. https://doi.org/10.1093/ref:odnb/4940.

Fitzmaurice, Susan. 2004. The meanings and uses of the progressive construction in an early eighteenth-century English network. In A. Curzan and K. Emmons (eds.). *Studies in the history of the English language II*, 131–174. Berlin: de Gruyter.

Fonteyn, Lauren and Andrea Nini. My alternation, my rules: Investigating syntactic variation in individual Englishes. *Cognitive Linguistics* (Special issue: *Constructionist Approaches to Individual Grammars*).

Gotti, Maurizio. 2013. The formation of the Royal Society as a community of practice and discourse. In J. Kopaczyk and A.H. Jucker (eds.). *Communities of practice in the history of English*, 269–285. Amsterdam/Philadelphia: John Benjamins.

Guy, Gregory and Sally Boyd. 1990. The development of a morphological Class. *Language Variation and Change* 2 (1): 1–18.

Hanson, Craig Ashley. 2009. *The English virtuoso: Art, medicine, and antiquarianism in the age of Empiricism*. Chicago, IL: University of Chicago Press.

Howard-Hill, T.H. 2006. Early modern printers and the standardization of English spelling. *The Modern Language Review* 101 (1): 16–29.

Kopaczyk, Joanna and Andreas H. Jucker (eds.). 2013. *Communities of practice in the history of English*. Amsterdam and Philadelphia: John Benjamins.

Kroch, Anthony, Beatrice Santorini and Lauren Delfs. 2004. *The Penn-Helsinki Parsed Corpus of Early Modern English* (PPCEME). University of Pennsylvania: Department of Linguistics. CD-ROM, first edn., release 3. www.ling.upenn.edu/ppche/ppche-release-2016/PPCEME-RELEASE-3.

Kytö, Merja and Terry Walker. 2006. *Guide to* A Corpus of English Dialogues 1560–1760 (Studia Anglistica Upsaliensia 130). Uppsala: Acta Universitatis Upsaliensis.

Labov, William. 2007. Transmission and diffusion. *Language* 83: 344–387.

Manjavacas, Enrique A. and Peter Petré. 2017. Enabling annotation of historical corpora in an asynchronous collaborative environment. In *Proceedings of DATeCH2017, Göttingen, Germany, June 01–02, 2017*, 6 pages. http://dx.doi.org/10.1145/3078081.3078089.

Milroy, James and Lesley Milroy. 1997. Network structure and linguistic change. In N. Coupland and A. Jaworski (eds.). *Sociolinguistics*, 199–211. London: Palgrave.

Milroy, Lesley and James Milroy. 1992. Social network and social class: Toward an integrated sociolinguistic model. *Language in Society* 21 (1): 1–26.

Nevalainen, Terttu, Helena Raumolin-Brunberg and Heikki Mannila. 2011. The diffusion of language change in real-time. *Language Variation and Change* 23: 1–43.

Nevalainen, Terttu. 2015. Social networks and language change in Tudor and Stuart London – only connect? *English Language and Linguistics* 19 (2): 269–292.

Nurmi, Arja, Ann Taylor, Anthony Warner, Susan Pintzuk and Terttu Nevalainen. 2006. *Parsed Corpus of Early English Correspondence, tagged version* (PCEEC). Compiled by the CEEC Project Team. York and Helsinki: University of York and University of Helsinki. Distributed through the Oxford Text Archive.

Petré, Peter and Freek Van de Velde. 2018. The real-time dynamics of the individual and the community in grammaticalization. *Language* 94 (4): 867–901.

Raumolin-Brunberg, Helena. 2009. Lifespan changes in the language of three early modern gentlemen. In A. Nurmi, M. Nevala and M. Palander-Collin (eds.). *The language of daily life in England (1400–1800)* (Pragmatics & Beyond 183), 165–196. Amsterdam: Benjamins.

Repo, Liina. 2018. Errors and corrections: Early Modern English errata lists in 1529–1700 and their connection to prescriptivism. Turku: Faculty of Humanities, MA thesis. http://www.utupub.fi/handle/10024/146176.

Rissanen, Matti, Merja Kytö, Leena Kahlas-Tarkka, Matti Kilpiö, Saara Nevanlinna, Irma Taavitsainen, Terttu Nevalainen and Helena Raumolin-Brunberg. 1991. *Helsinki Corpus of English Texts.* Department of Modern Languages: University of Helsinki.

Rivers, Isabel. 2004. Tillotson, John (1630–1694), archbishop of Canterbury. *Oxford dictionary of national biography.* Oxford: Oxford University Press. https://doi-org/10.1093/ref:odnb/27449.

Sairio, Anni. 2009. Methodological and practical aspects of historical network analysis. In A. Nurmi, M. Nevala and M. Palander-Collin (eds.). *The language of daily life in England (1400–1800)* (Pragmatics & Beyond 183), 107–135. Amsterdam: Benjamins.

Sankoff, Gillian. 2005. Cross-sectional and longitudinal studies in sociolinguistics. In P. Trudgill (ed.). *Sociolinguistics: An international handbook of the science of language and society*, 1003–1013. Berlin: De Gruyter Mouton

Schmid, Hans-Jörg. (Forthcoming). *The dynamics of the linguistic system: Usage, conventionalization, and entrenchment.* Oxford: Oxford University Press.

Standing, William and Peter Petré. (Submitted). Lifespan change versus intergenerational incrementation in the schematization of syntactic constructions. In I. Buchstaller, S. Wagner and K. Beaman (eds.). *Panel studies of variation and change, vol. II*. Oxford: Routledge.

Standing, William, Odile A.O. Strik and Peter Petré. (Submitted). Change versus stability in syntactic constructions of Early Modern English networked individuals. *Journal of English Linguistics* (Special issue: *The Role of an Individual Speaker in Linguistic Change*).

Steels, Luc. 2000. Language as a complex adaptive system. In M. Schoenauer, K. Deb, G. Rudolph, X. Yao, E. Lutton, J.J. Merelo and H-P. Schwefel (eds.). *Parallel Problem Solving from Nature (PPSN) VI* (Lecture Notes in Computer Science 1917), 17–26. New York: Springer.

Taavitsainen, Irma, Päivi Pahta, Turo Hiltunen, Martti Mäkinen, Ville Marttila, Maura Ratia, Carla Suhr and Jukka Tyrkkö. 2010. *Early Modern English Medical Texts* (EMEMT). CD-ROM. Amsterdam: John Benjamins.

Theobald, Martin, Jonathan Siddharth and Andreas Paepcke. 2008. SpotSigs: Robust and efficient near duplicate detection in large web collections. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 563–570. New York: ACM. https://dl.acm.org/citation.cfm?id=1390431&dl=ACM&coll=DL.

Trudgill, Peter. 2011. *Sociolinguistic typology: Social determinants of linguistic complexity.* Oxford: Oxford University Press.

Van de Velde, Freek and Peter Petré. 2017. Linking grammaticalization to historical demography. Paper presented at *Historical Sociolinguistics Network*, New York, April 6–7.

Van de Velde, Freek. 2014. Degeneracy: The maintenance of constructional networks. In R. Boogaart, T. Colleman and G. Rutten (eds.). *Extending the scope of construction grammar*, 141179. Berlin: De Gruyter Mouton.

Wagner, Suzanne Evans. 2012. Age grading in sociolinguistic theory. *Language and Linguistics Compass* 6 (6): 371–382.

Walker, Terry. 2017. "he saith yt he thinkes yt": Linguistic factors influencing third person singular present tense verb inflection in Early Modern English depositions. *Studia Neophilologica* 89 (1): 133–346.

Yáñez-Bouza, Nuria. 2011. ARCHER past and present (1990–2010). *ICAME Journal* 35: 205–236.