# The EMMA corpus (release 1.0)
# *Early Modern Multiloquent Authors*

*Notice*: This manual is intended as a practical starting point for using EMMA. We are currently preparing a more comprehensive article with a detailed description of EMMA and the opportunities it offers, to be submitted to the ICAME journal. Thank you for referring to the article manuscript in publications in which EMMA features:

Petré, Peter; Lynn Anthonissen; Sara Budts; Enrique Manjavacas; William Standing; and Odile A.O. Strik. 2019. Early-Modern Multiloquent Authors (EMMA): Designing a large-scale corpus of individuals' languages. *ICAME journal*, manuscript.

## 1. Background

EMMA (*Early Modern Multiloquent Authors*) is a sample of 50 of the most prolific English writers born in the 17th century, who mostly belonged to the London-based elite. 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. With the help of the corpus, the project wants to settle how much 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 for all kinds of linguistic research that benefits from the individual perspective.

| | |
|---|---|
| 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, Diane Koek, BA & MA students from the University of Antwerp |
| Time of compilation | 2015–present |
| Size | 90 million words (inclusive non-English text); 88.5 million (English only) |
| Language | English |
| Number of texts/samples | 13750 |
| Period | 1623–1757 |
| Released | 2018 |
| Funding | H2020 - European Research Council (ERC) (Project ID 639008) |
| Project home page | www.uantwerpen.be/mind-bending-grammars |

## 2. Availability

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 free release includes the part of EMMA that is in the public domain. The remainder of EMMA will be made available once the source texts have entered the public domain (around 2021). Researchers from universities with a subscription to the source database EEBO-TCP Phase II already get access to the full corpus.

## 3. Technical information

The body of texts was mainly collected from the EEBO[1] and ECCO[2] databases following an extensive author selection (and for ECCO data, an OCR correction[3]) process. The corpus is tokenized and encoded in Unicode UTF-8. It comes in XML and plain TXT formats.

The open source software CosyCat[4] (Collaborative Synchronized Corpus Analysis Toolkit) has been developed for querying and annotating the corpus (currently in alpha). CosyCat queries a version of EMMA that is indexed by BlackLab[5]. Rich metadata for the corpus is stored as xml-headers.

## 4. Structure and selection criteria

The EMMA corpus is a large-scale specialized corpus that comprises the writings of 50 carefully selected authors across 5 generations.

For the author selection, we drafted a number of criteria. On the individual level we looked for: (i) a large body of work comprising at least 500,000 words; (ii) a relatively even distribution of works across a long career; (iii) a demonstrable link to London society; (iv) further social, political, and stylistic connections to other individuals in the selection. 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 these criteria. On the level of the author selection as a whole, we valued a distribution across different main genres such as religious writing, science, drama, and letter writing. In addition, we included a few authors who were not strongly connected socially to London and/or the other individuals, but who otherwise fulfilled individual criteria. These authors may serve as a control group when looking at the spread of linguistic changes through the social networks of the time. Table 1 gives an overview of the authors in the EMMA corpus; Figure 1 presents the corpus distribution.

| | Id | Author | Description | Born | Died |
|---|---|---|---|---|---|
| Generation 1 | 101 | Heylyn, Peter | churchman, author | 1599 | 1662 |
| | 102 | Prynne, William | lawyer, author, political figure | 1600 | 1669 |
| | 103 | Davenant, Sir William | playwright | 1606 | 1668 |
| | 104 | Fuller, Thomas | churchman, historian | 1607 | 1661 |
| | 105 | Milton, John | poet | 1608 | 1674 |
| | 106 | Taylor, Jeremy | cleric, author | 1613 | 1667 |
| | 107 | More, Henry | philosopher | 1614 | 1687 |
| | 109 | Baxter, Richard | church leader, poet, theologian | 1615 | 1691 |
| | 110 | Owen, John | church leader, theologian | 1616 | 1683 |
| | 111 | L'Estrange, Roger | pamphleteer, author, politician, Licenser of the Press | 1616 | 1704 |
| Generation 2 | 201 | Boyle, Roger | soldier, dramatist, politician | 1621 | 1679 |
| | 202 | Pierce, Thomas | churchman | 1622 | 1691 |
| | 204 | Fox, George | Quaker founder | 1624 | 1691 |
| | 205 | Boyle, Robert | natural philosopher, chemist, physicist, inventor | 1627 | 1691 |
| | 206 | Swinnock, George | churchman | 1627 | 1673 |
| | 207 | Bunyan, John | writer, preacher | 1628 | 1688 |
| | 208 | Flavell, John | clergyman, author | 1630 | 1691 |
| | 209 | Tillotson, John | Archbishop of Canterbury | 1630 | 1694 |
| | 210 | Dryden, John | poet, playwright, critic, translator | 1631 | 1700 |
| | 211 | Cavendish, Margaret | philosopher, poet, scientist, fiction-writer, playwright | 1623 | 1673 |
| | 215 | Phillips, John | author, translator, secretary to Milton | 1631 | 1706 |

[1] Early English Books Online (eebo.chadwyck.com)

[2] Eighteenth Century Collections Online (quod.lib.umich.edu/e/ecco)

[3] OCR-ed ECCO texts were manually corrected with the correction tool provided by 18th Connect (18thconnect.org). XML tags such as highlight tags, quote tags, note tags and form work tags were inserted as well.

[4] github.com/emanjavacas/cosycat

[5] github.com/INL/BlackLab

| | | | | | |
|---|---|---|---|---|---|
| Generation 3 | 301 | Stillingfleet, Edward | theologian, scholar | 1635 | 1699 |
| | 302 | Whitehead, George | Quaker leader | 1637 | 1724 |
| | 303 | Whitby, Daniel | theologian, biblical commentator | 1638 | 1726 |
| | 305 | Mather, Increase | puritan minister, colonist | 1639 | 1723 |
| | 306 | Sherlock, William | church leader | 1641 | 1701 |
| | 307 | Keach, Benjamin | preacher | 1640 | 1704 |
| | 308 | Crouch, Nathaniel | printer, bookseller, historian | 1640 | 1725 |
| | 310 | Behn, Aphra | playwright, poet, translator, author, spy | 1640 | 1689 |
| | 311 | Crowne, John | dramatist | 1641 | 1712 |
| | 312 | Burnet, Gilbert | philosopher, historian, bishop | 1643 | 1715 |
| | 313 | Salmon, William | doctor | 1644 | 1713 |
| | 314 | Penn, William | Quaker, founder of Pennsylvania | 1644 | 1718 |
| Generation 4 | 401 | D'Urfey, Thomas | writer, poet | 1653 | 1723 |
| | 402 | Wake, William | Archbishop of Canterbury | 1657 | 1737 |
| | 403 | Dennis, John | playwright | 1657 | 1734 |
| | 404 | Dunton, John | bookseller, author, publisher | 1659 | 1733 |
| | 405 | Defoe, Daniel | author, journalist, spy | 1660 | 1731 |
| | 406 | Mather, Cotton | minister, author, pamphleteer | 1663 | 1728 |
| | 407 | Harris, John | writer, scientist, priest | 1666 | 1719 |
| | 408 | Swift, Jonathan | author, poet, satirist, pamphleteer, cleric | 1667 | 1745 |
| | 409 | Whiston, William | theologian, historian, mathematician | 1667 | 1752 |
| | 410 | Ward, 'Ned', Edward | satirist, publican | 1667 | 1731 |
| Generation 5 | 501 | Cibber, Colley | playwright, actor, manager, Poet Laureate | 1671 | 1757 |
| | 502 | Steele, Richard | writer, politician | 1672 | 1729 |
| | 503 | Addison, Joseph | essayist, poet, playwright, politician | 1672 | 1719 |
| | 504 | Oldmixon, John | historian, author | 1673 | 1742 |
| | 505 | Clarke, Samuel | philosopher, clergyman | 1675 | 1729 |
| | 506 | Hoadly, Benjamin | clergyman, bishop | 1676 | 1761 |
| | 508 | Jacob, Giles | author, legal writer | 1686 | 1744 |

Table 1. Authors in the EMMA corpus

In addition, a more extensive author metadata database is underway with information, which, *inter alia*, includes quantifiable social network information and mobility history of each author. The metadata database is currently in alpha. An official release wt

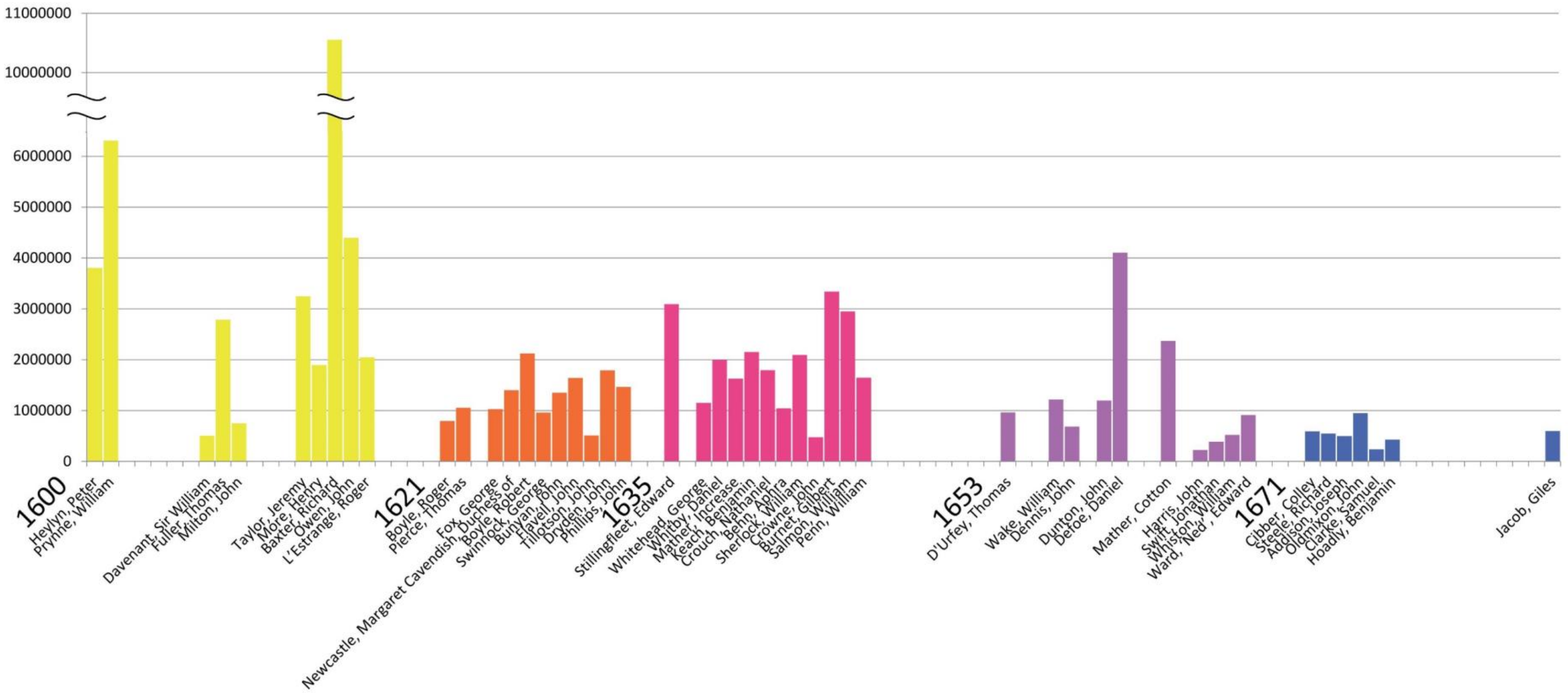


Fig. 1 Distribution of the EMMA corpus

Genre balance was not a primary criterion. However, the corpus contains considerable amounts of text from the predominant written genres of the 17th century. The following is a table of those genres that are represented by at least 50,000 words in every generation. More information on the genre classification can be found below (Section 5.3).

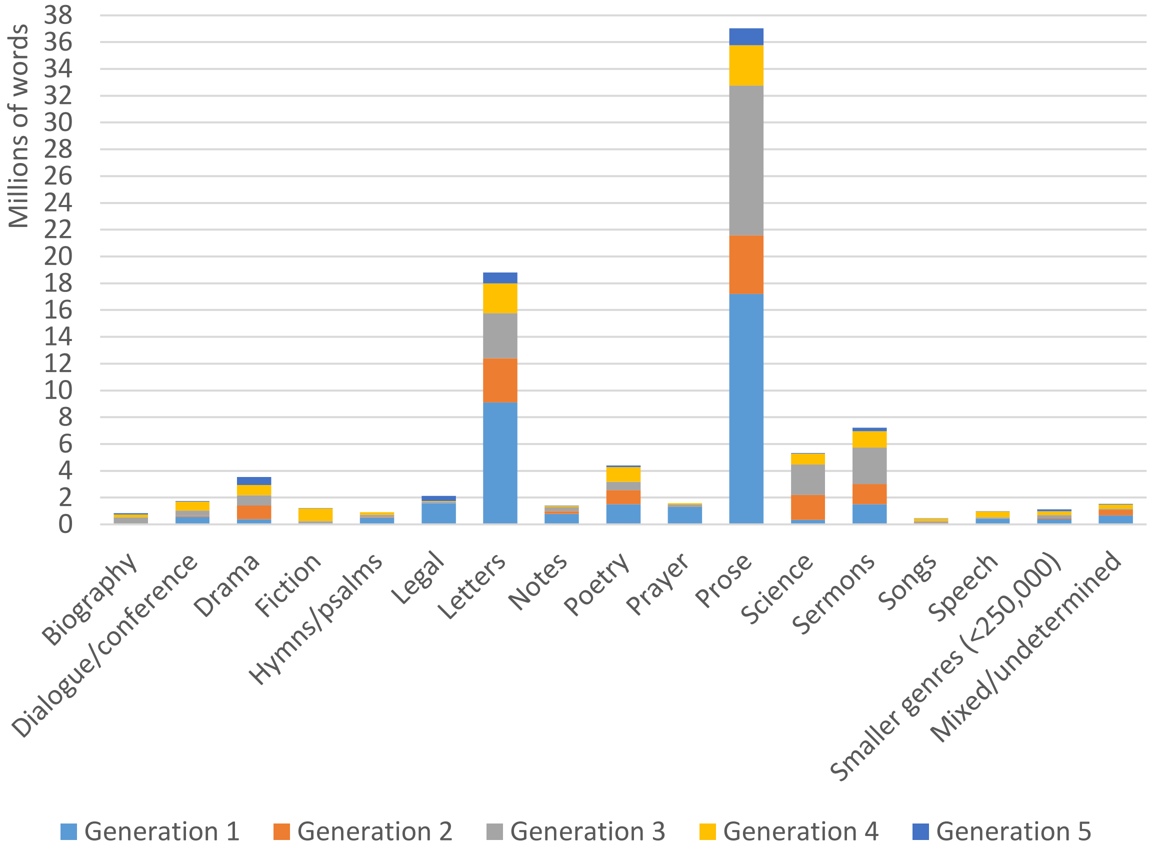


Table 2. Genre distribution

## 5. Metadata

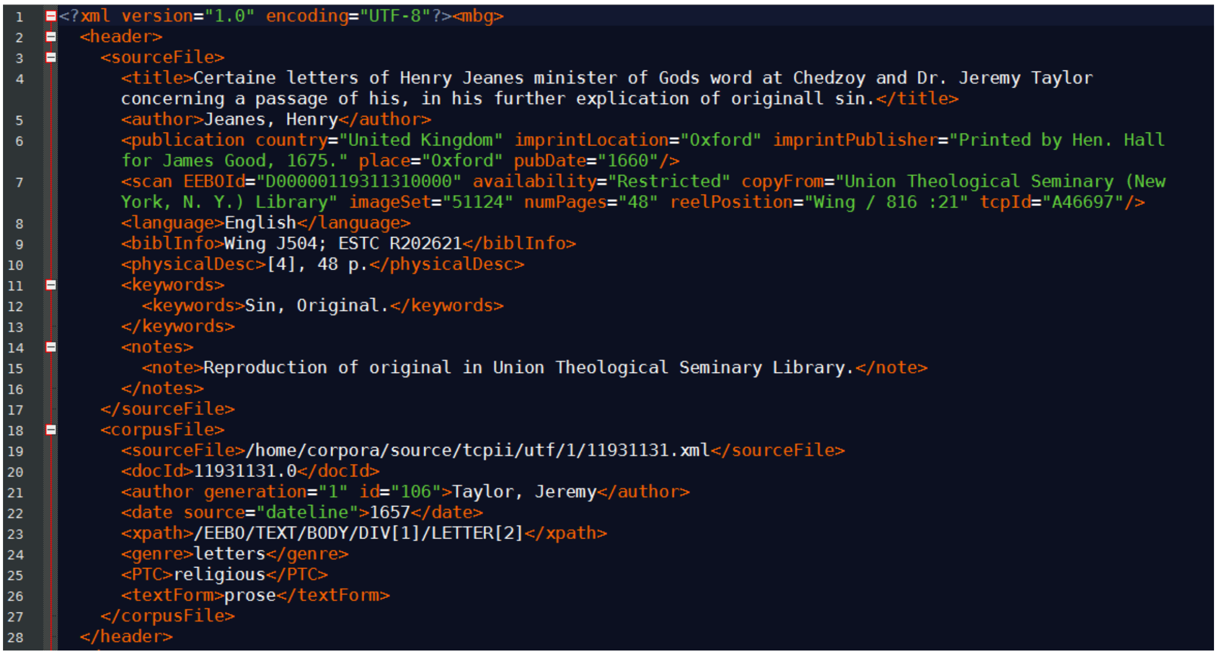
### 5.1 Metadata in the Excel sheet

The metadata Excel 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. 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 Phase II can already download the complete EMMA corpus; an open access version without EEBO Phase II is available for those without subscription. Apart from the word counts, text-specific metadata are stored in the XML **<header>** element in EMMA's corpus files.

### 5.2 XML Headers

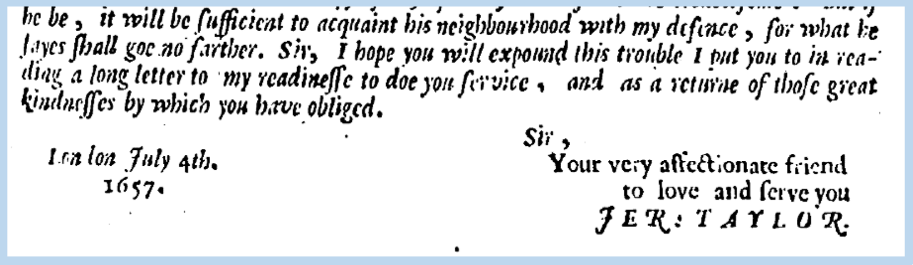
Text-specific metadata are stored in the XML headers. Most of the information was automatically retrieved from the EEBO and ECCO databases and is retained under the **<sourceFile>** element. However, we have spent a great deal of time verifying the metadata, especially date and authorship, and have also 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 is attached to the header under **<corpusFile>**.

Fig. 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 Fig. 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
<?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>religious</PTC>
      <textForm>prose</textForm>
    </corpusFile>
  </header>
```

Fig. 2 XML header



he 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.

Sir,

London July 4th.
1657.

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

Fig. 3 Metadata verification

### 5.3 Genre classification

Another type of contextual enrichment is genre classification. Genre balance in itself was not a primary criterion, but the corpus does contain considerable amounts of text from the predominant written genres of the 17th century. The current classification is inspired by the systems used in the ARCHER and Helsinki corpora, and has been improved by an automatic genre classification tool (developed by Arthur Nieuwland during an AI internship at *Mind-Bending Grammars*). The classification is still preliminary, and at times remains underspecified. Further revision is planned for a future release.

*Genre classification on three levels*

| CLASSIFICATION | XML HEADER NAME | PRACTICE |
|---|---|---|
| Text form | textForm | General distinction between:<br>– prose<br>– verse |
| Prototypical text category | PTC | General distinction between:<br>– imaginative<br>– non-imaginative<br>– 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. |

*Prototypical text category*

| PROTOTYPICAL TEXT CATEGORY | [text types] |
|---|---|
| imaginative<br>~fiction | Fiction, romance, drama, poetry, etc. |
| nonimaginative<br>~non-fiction | Nonimaginative narratives and descriptive and/or argumentative texts on non-religious matters<br>e.g. history, biography, memoirs, treatise, essay, document, law, handbook, science, philosophy, education, personal correspondence (non-religious), diary, description of foreign countries, etc. |
| religious texts | Religious instruction, e.g. treatise, essay homily, rule, sermon, catechism etc.<br>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:<br>– if the text contains **various text types**, e.g. a letter and a poem<br>– in the case of **biographies/histories** of religious persons/institutions, which are labelled non-imaginative+religious<br>– 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 |

## *Genre*

| GENRE<br>[label] | [description] | [subcategories] |
|---|---|---|
| prose | generic label for argumentative and/or descriptive prose, not part of any of the other ‘prose’ text categories | large prose subcategories (sermons, legal, scientific texts) have a separate genre label |
| science | scientific texts | science<br>science_chemistry<br>science_geography<br>science_mathematics<br>science_medicine<br>science_physics |
| legal | legal texts | legal |
| letters | letters | letters<br>letters_monitory<br>letters_pastoral (= pastoral letters and charges) |
| sermons | orations or lectures by a member of the clergy | sermons_election<br>sermons_execution<br>sermons_fast-day<br>sermons_funeral |
| satire | satires | satire |
| fiction | imaginative narrative prose | fiction |
| drama | | drama<br>drama_comedy<br>drama_farce<br>drama_masque<br>drama_opera<br>drama_prologue/epilogue<br>drama_tragedy<br>drama_tragicomedy |
| poetry | poetical work | poetry<br>poetry_burlesque<br>poetry_elegy<br>poetry_epic<br>poetry_epigram<br>poetry_heroic<br>poetry_miscellany (various types of poems)<br>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 |
| v | various other minority genre categories | v_advertisement<br>v_fable<br>v_parable<br>v_testimony<br>v_prayer |
| miscellany | mix of different text types | miscellany |
| undetermined | genre is unclear | undetermined |

## 6. Foreign language tagging

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 single clauses 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. These estimates are included in the metadata file, where they are subtracted from the raw word count.

## 7. Work in progress

Currently there are two ongoing projects that further enrich the corpus and its metadata. The results of these projects will be integrated in a future release of EMMA.

First, the genre revision is still ongoing. Information from the genre classification that suits our corpus data better than the divisions made in the Helsinki Corpus. EMMA's current genre classification is based on explicit genre indicators in the title or previously assigned tags. This typology has received some corrections from a bottom-up, AI-based approach that classifies the texts based on their content rather than their metadata, but a systematic integration of both systems has not yet been carried out.

The second ongoing project is concerned with spelling normalization. The normalization is carried out with the University of Lancaster's VARD-tool[6], 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. The spelling normalization process is entirely reversible, as the original forms will be preserved as attributes of the word tags.

## 8. Contact

Peter Petré: peter.petre@uantwerpen.be
EMMA: emma@uantwerpen.be
Project website: www.uantwerpen.be/en/projects/mind-bending-grammars
CosyCat: github.com/emanjavacas/cosycat

[6] ucrel.lancs.ac.uk/vard/about/