Emerging Voices Corpus, version 1.0
User's manual

George Walkden

September 2019

Contents

1 Introduction

The Emerging Voices Corpus is a small corpus (47,481 words; 53,567 tokens including punctuation) of Early Modern English (1500–1800). You can cite it as follows:

> Walkden, George. 2019. The Emerging Voices Corpus, version 1.0. http://walkden.space/voices/

2 Corpus composition and text selection

The corpus consists of 24 text samples of 2,000 words (+/- 200), each by a different author:

- 16th century: 8 authors (5 female, 3 male)
- 17th century: 8 authors (6 female, 2 male)
- 18th century: 8 authors (4 female, 4 male)

The texts are *egodocuments*:

> 'those historical sources in which the researcher is faced with an 'I' ... as the writing and describing subject with a continuous presence in the text' (Presser, 1958)

> 'a broad category comprising several forms of autobiographical texts, including autobiographies, memoirs, diaries, travel journals, and personal letters' (Mascuch et al., 2016)

The corpus is diachronically balanced and relatively genre-homogeneous, but it is *not* intended as representative of any particular population. Rather, efforts have been made the deliberately *over*represent (at least as far as the historical record is concerned) writers who are underrepresented in existing corpora: women, writers from outside England, people of colour, those not from the upper classes.

Texts were selected on the basis of being freely accessible on the web and in the public domain, following the criteria mentioned above; over half of the texts are taken from archive.org. A spreadsheet provided with the corpus gives additional metadata and information about the authors and texts included.

3 Correction and annotation

The corpus was constructed as part of a course on Early Modern English taught at Masters level at the University of Konstanz in the summer semester of 2019, in which 23 students participated (see Acknowledgements). As part of the assessment for the course, each student had to do the following:

1. Choose (or be assigned) an author to 'adopt'
2. Prepare the text for automatic tagging and lemmatization (correct OCR errors, separate out clitics and punctuation, etc.)
3. Correct the automatic tagging and lemmatization

The students' work was checked and corrected where necessary by me after steps 2 and 3. Spelling normalization was carried out after step 2 using VARD 2.5.4, and the output of VARD was tagged and lemmatized using TreeTagger and the Penn Treebank tagset.

In-text metadata, enclosed within angle brackets, is of two types:

1. Sections (e.g. in cases where several letters are combined to produce a single 'text'), for instance `<section_1>`.
2. VARD-normalized words, for instance

```
<normalised orig="Quin" auto="true">
Queen NN queen
</normalised>
```

4 Known issues and intended uses

The corpus is intended to be used for morphological, syntactic, semantic, pragmatic and lexical research into varieties of Early Modern English. It is not suitable for orthographic or phonological research, due to the variety of sources used (printed books, transcripts of letters, etc.), though we have tried to preserve the original orthography of each source as faithfully as possible. It is also obviously not suitable for studies of the book (or text) as physical object: marginalia and words duplicated at the start and end of pages have been removed, for instance. Finally, due to its small size, the corpus may not yield reliable results when rare phenomena are the object of investigation.

5 Acknowledgements

Thanks first and foremost to all the students who helped to construct this corpus: Andreas Choumas, Sylvia Demasi, Michael Diwersy, Rike Grotjahn, Jana Hägele, Ludmilla Hartmann, Leah Illi, Julia Keller, Diana Kempel, Zsofia Lelner, Selina Lüttin, Alisa Mader, Maria Pfleger, Sven Reppenhagen, Wibke Rhein, Lisa Rieger, Fabian Schneider, Lukas Sedlmeier, Nico Stoll, Sven Timmer, Tjarko van der Linde, Lena Westhauser, and Selina Winkler. Thanks also to Michael Walkden, who advised me on text selection.

References

Mascuch, M., Dekker, R., and Baggerman, A. (2016). Egodocuments and history: a short account of the *longue durée*. *The Historian*, 78:11–56.

Presser, J. (1958). Memoires als geschiedbron. In *Algemene Winkler Prins Encyclopedie*, volume 7, pages 208–210. Elsevier, Amsterdam.