**Title: A Computational Approach to Analysing the Corpus of Oz Early English**

Author:

Martin Schweinberger\*

The University of Queensland

Address: School of Languages and Cultures, Gordon Greenwood Building, Union Road,

The University of Queensland, St Lucia, QLD 4072, Australia

Phone: +61 7 3365-6374

Email: [m.schweinberger@uq.edu.au](mailto:m.schweinberger@uq.edu.au)

ORCID: 0000-0003-1923-9153

\*corresponding author

**Abstract**

This study analyses the content of the private written section of the *Corpus of Oz Early English* (COOEE) using quantitative text analytic methods. The private writing section of the COOEE consists of 607 unpublished letters and diary entries written in Australia or by native Australians on travels, between 1788 and 1900. This study focuses on keywords that are used and the topics in this corpus. As such, the present study sets out to exemplify how computational methods such as keyword extraction, network analysis, and topic modelling can assist historical linguists, dialectologists, and corpus linguists in unearthing patterns and topics that would be hard to identify using traditional methods. Keywords were extracted by comparing the frequencies of words in the COOEE against their frequency in a corpus consisting of letters written in Britain by British authors roughly during the same time period. The keywords show that authors represented in the COOEE mostly write about their physical surroundings, their living situation, and life in Down Under. A latent dirichlet allocation-based topic model reports that six topics figure prominent in the texts ranging from private family issues, over exploration and the landscape as well as encounters with the indigenous people of Australia to employment options, their journey to Australia, and circumstances after their arrival. The results presented here exemplify how dialectology, historical linguistics, and studies describing the content of collections of electronic texts can profit from adopting computational methods developed in natural language processing.

**Keywords**

Corpus of Oz Early English (COOEE), Text Mining, Corpus Linguistics, Keyword extraction, Network Analysis, Topic Model

**Author bionotes**

Martin Schweinberger is Lecturer in Applied Linguistics in the School of Languages and Cultures at the University of Queensland, Director of the Language Technology and Data Analysis Laboratory, and Associate Professor II in the AcqVA Aurora Center at the Artic University of Norway, Tromsø. His research focuses on statistical modelling and visualization of linguistic data, language variation and change, and acquisition of and variability in language use.

1. Introduction

While the use of corpora in linguistics has been increasing dramatically ever since they became a viable tool for analysing natural language in the 1980s (see Lindquist 2009), analyses of corpus data almost exclusively focus on the variability in the occurrence of specific linguistic features. Rather than focusing on linguistic aspects of the corpus, this paper explores the semantic and narrative content of a corpus, the *Corpus of Oz Early English* (COOEE) (Fritz 2004). To this end, the present study uses the private written register of the COOEE which consists of 607 unpublished letters and diary entries written in Australia or by native Australians on travels between 1788 and 1900. Specifically, the present study focuses on what sets the contents of this corpus apart from a parallel corpus of letters written in Great Britain compiled at the John Rylands University Library of Manchester, *The English language of the north-west in the late Modern English period: A Corpus of late 18th century Prose* (CoRD, Denison & van Bergen 2003).

In addition to exploring and analysing the semantic and narrative content of the COOEE, the study sets out to exemplify how and to what extent the use of quantitative text analytic methods, such as keyword extraction and topic modelling, can assist in summarizing and exploring the lexical and semantic content of large amounts of texts as provided in corpora or other collections of textual data. This approach is not only innovative as it uses an inter-disciplinary approach but it is also increasingly important to inspect the content of large amounts of texts due to the ever-increasing availability and use of Big Data in the humanities.

The next section presents previous research on amplifiers and changes in amplifier systems. Section 3 provides information about the corpus data used in the current study, elaborates on the steps undertaken during data processing, and describes the statistical analyses that were applied to the data. Section 4 presents the results of the statistical analysis while section 5 discusses the results in light of previous research and evaluates shortcomings of the present analysis.

As such, the present study addresses the following research questions:

RQ 1: What concepts are written more about in the COOEE compared to similar private letters correspondence written in Great Britain?

RQ 2: Can we observe changes in the key concepts over time?

RQ 3: What topics are written about in the COOEE?

RQ 4: Can we observe changes in the prominence of these topics over time?

2. Data and Methodology

The following section consists of two subsections: the first subsection provides information about the COOEE as well as the control corpus, the CoRD, and describes the data processing while the second subsection describes the statistical methods that have been applied to the data.

2.1 Corpus Description and Data Processing

The entire COOEE comprises 1,353 unpublished letters, published books, and historical texts written in Australia, New Zealand or Norfolk Island, or by native Australians on travels, between 1788 and 1900. In particular, the COOEE represents four registers: the Speech-based Register (SB), the Private Written Register (PrW), the Public Written Register (PcW) and the register of Government English (GE). However, as this study aims to analyse topics emigrants and early settlers of Australia talked about in their private communication, the study only focuses on the PrW register which represents personal letters and diary entries in which authors confided their private joys and sorrows. This PrW register consists of 607 private letters amounting to 329,233 tokens (172,912types) (see Table 1)[[1]](#footnote-1). We decided against using the periodization proposed by the COOEE corpus compiler and work with a division of the corpus data into six approximately equal periods instead (see below).

Of these 607 letters, 178 were composed by women amounting to 29.3 percent. Letters were on average about 542,4 words long. Due to the semi-literate status of the writers, orthography in the COOEE is variable and occasionally mirrors the writers’ pronunciation.

Table 1: Overview of the number of letters, types and tokens per period in the COOEE.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Period** | **Author Gender** | **Letters (N)** | **Types (N)** | **Tokens (N)** |
| 1788-1800 | female | 9 | 2,799 | 4,010 |
| 1788-1800 | male | 19 | 11,532 | 24,225 |
| 1801-1820 | female | 20 | 6,595 | 13,914 |
| 1801-1820 | male | 23 | 7,515 | 12,773 |
| 1821-1840 | female | 57 | 21,153 | 32,605 |
| 1821-1840 | male | 52 | 25,209 | 48,851 |
| 1841-1860 | female | 56 | 19,924 | 33,302 |
| 1841-1860 | male | 61 | 19,150 | 29,207 |
| 1861-1880 | female | 19 | 4,715 | 6,964 |
| 1861-1880 | male | 41 | 17,996 | 45,283 |
| 1881-1900 | female | 17 | 3,813 | 7,623 |
| 1881-1900 | male | 233 | 32,511 | 70,476 |
| **Total** |  | **607** | **172,912** | **329,233** |

Regarding the geographic location of authors, the south-east of Australia is substantively overrepresented (see Figure 1).

|  |
| --- |
| Figure 1: Geographic location of authors in COOEE (9 letters missing as they were written outside of Australia: 6 at sea, 1 in Italy, 3 in Great Britain). |

In order to find and extract keywords from the COOEE, we compared the frequencies of words in the COOEE to a control corpus which consisted of letters written in Great Britain. The control corpus, *The English language of the north-west in the late Modern English period: A Corpus of late 18th century Prose* (CoRD, Denison & van Bergen 2003) was compiled at the John Rylands University Library of Manchester (see Table 2 for an overview).

Table 2: Overview of the number of letters, types and tokens per period in the CoRD – the control corpus.

|  |  |  |  |
| --- | --- | --- | --- |
| **Period** | **Letters (N)** | **Types (N)** | **Tokens (N)** |
| 1788-1800 | 1,790 | 190,871 | 279,218 |
| 1881-1900 | 2 | 172 | 215 |
| unknown | 35 | 2,777 | 3,670 |
| **Total** |  | **193,820** | **283,103** |

While the letters in the control corpus were written mostly during an early and relatively short period (1788 to 1800), the data lends itself to serve as a control because it also represents private letters.

The data within each corpus was processed in R (R Core Team 2021). The data were loaded and annotated with metadata (sociodemographic information about the author and information relating to the data itself such as when and where the letter was written). The resulting table was then saved for further processing. For the topic modelling, the letters were split into sentences resulting in a table holding 13,931 sentences which formed the basis for the Latent Dirichlet Allocation which represents the core procedure for the topic modelling.

The textual data was then cleaned by removing so-called stops words, i.e., words not carrying semantic but rather indexical or grammatical information such as pronouns, determiners, and adverbs (e.g., *the*, *a*, *an*, *that*, *many*, *his*, *she*, *it*, etc.). For keyword extraction, any elements consisting of non-alphanumeric characters were also removed – including punctuation and quotation marks. In addition, the data were manually cross-checked to minimize the number of errors arising from the (semi-) automatic data processing.

2.2 Statistical Methods

The present study makes use of two types of analyses:

(1) keyword analysis which is applied both to the COOEE as a whole to extract words that are significantly overused in the COOEE compared to a control corpus and within the COOEE to extract words that are significantly overused in certain periods the COOEE;

(2) Latent dirichlet allocation-based topic modelling (Blei et al., 2003) which is used to find thematically coherent topics or themes within the COOEE. The results of the topic model also allow us to determine, for each sentence, what topic a sentence is associated with, which in turn enables us to analyse changes in the prominence of topics across periods.

2.2.1 Keyword Analysis

The keyword analysis used Fishers’ Exact test to determine for every word if it occurred significantly more frequently in the COOEE compared to the CoRD while controlling for all other words in the respective corpora. As this resulted in a large number of tests, we used Bonferroni corrections to avoid an inflation of alpha-error (false positive results) (see Field, Miles, & Field 2012: 428-432). If a word occurred significantly more frequently in the COOEE, it was regarded as a keyword. As this procedure detected a very high number of keywords (2,591), we decided to further differentiate between keywords with a small effect size (phi <= .002) and keywords with a meaningful effect size (phi > .002) – only the latter category of keywords (24) was displayed in the network graph which was used to visualize the co-occurrence of keywords in the COOEE (the network co-occurrence graph was generated using the quanteda package in R, see Benoit et al. 2018).

To tap into thematic shifts, keywords were also extracted for individual periods by comparing the frequencies of words in a given period against the frequency of this word in other periods. If a word occurred significantly more frequently in a given period, it was deemed a keyword. Again, Bonferroni corrections were applied to safeguard against alpha-error inflation.

2.2.2 Topic Modelling

To assess thematically coherent topics and to analyse shifts in the prominence of these topics, we applied latent dirichlet allocation-based topic modelling (Blei, Ng, and Jordan 2003) using the seededlda package in R (Watanabe and Xuan-Hieu 2021). In contrast to the keyword analysis, the data was stemmed using the stemDocument function from the tm package in R (see Feinerer and Hornik 2020) and split into sentences to find groups of words that co-occurred in the same sentences and thus correlated in use (these groups of correlating terms represent the topics). The topic modelling consisted of two steps: a first, unsupervised data-driven topic modelling process which served to identify seeding terms which were then used in a second step which consisted of a semi-supervised seeded topic modelling which aimed to find topics based on the seed-terms provided by the first step. The results of the topic modelling were manually inspected for models for 5 to 15 topics and it was determined, based on topic-indicative keywords, if the models provided meaningful results. The model which identified six topics provided the best results in that it resulted in six thematically coherent, non-overlapping topics with one additional, less coherent topic (other) that served as a bin category for sentences that could not be classified into any of the topics with sufficient confidence. The following section presents the results of the quantitative analysis.

3. Results

The following section reports on the findings of the study of the Private Written Register (PrW) of the COOEE consisting of 607 personal letters and diary entries written between 1788 and 1900. Before presenting the results of the LDA-based topic modelling, we will inspect the results of the keyword analysis.

3.1 Keyword Analysis

The keyword analysis detected 2,560 words that occurred significantly more frequently in the COOEE compared to the CoRD. This is a remarkably high number of keywords given the comparatively moderate size of the corpora and it indicates that the content of the COOEE is notably different from the CoRD – despite the similarity in register.

Due to the high number of significant keywords, only keywords with an effect size of phi > .02 are displayed in the following tables and visualizations. Table 3 provides on overview of these 22 keywords.

Table 3: Overview of words significantly overused in COOEE (keywords) with phi > .02 (COOEE total = 259,912, Control total = 119,501)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Word** | **COOEE** | **CoRD (control)** | **Χ2** | **φ (phi)** | **Significance (Boferroni corr.)** |
| miles | 2033 | 4 | 928.5 | 0.0495 | p<.001 |
| water | 1521 | 87 | 508.1 | 0.0366 | p<.001 |
| country | 1246 | 48 | 463.4 | 0.0349 | p<.001 |
| creek | 610 | 0 | 279.5 | 0.0271 | p<.001 |
| river | 666 | 14 | 272.2 | 0.0268 | p<.001 |
| camp | 593 | 1 | 269.2 | 0.0266 | p<.001 |
| natives | 554 | 0 | 253.6 | 0.0259 | p<.001 |
| south | 518 | 2 | 232.2 | 0.0247 | p<.001 |
| north | 520 | 7 | 221.2 | 0.0241 | p<.001 |
| found | 918 | 102 | 218 | 0.024 | p<.001 |
| course | 627 | 33 | 213.9 | 0.0237 | p<.001 |
| camels | 428 | 0 | 195.5 | 0.0227 | p<.001 |
| like | 735 | 69 | 195 | 0.0227 | p<.001 |
| wind | 421 | 5 | 180.3 | 0.0218 | p<.001 |
| east | 407 | 3 | 178.6 | 0.0217 | p<.001 |
| west | 405 | 5 | 173 | 0.0214 | p<.001 |
| range | 377 | 0 | 172 | 0.0213 | p<.001 |
| went | 716 | 79 | 170.6 | 0.0212 | p<.001 |
| feet | 413 | 9 | 167.5 | 0.021 | p<.001 |
| sydney | 348 | 0 | 158.7 | 0.0205 | p<.001 |
| large | 550 | 46 | 155.3 | 0.0202 | p<.001 |
| started | 353 | 3 | 153.8 | 0.0201 | p<.001 |

The keywords listed in table 3 almost exclusively denote physical attributes or refer to landmarks and landscapes.

The network co-occurrence graph below (see Figure 2) shows the co-occurrence of keywords within the COOEE to enable the assessment of what keywords are used by themselves and which keywords play a more central role in the letters and diary entries. According to the network analysis, keywords like *country*, *camp*, *creek*, *camels*, *course*, *east*, or *miles* that are likely to occur in context of exploration and travel within Australia are more central while terms like *ship*, *wind*, *like*, *rain*, or *nearly* that are more likely to occur in the context of journeys to Australia (rather than within Australia).

|  |
| --- |
| Figure 2: Network co-occurrence graph of keywords in the COOEE. |

Figure 3 shows the distribution of selected keywords across periods to give an impression of changes in the prominence of keywords.

|  |
| --- |
| Figure 3: Relative frequency of selected keywords across periods in the COOEE. |

Figure 3 shows that *australia*, *family*, and *travelled* tend to occur more towards the later periods while *convicts* occurred more in early writings. Terms like *colony*, *natives*, and *ship* do not exhibit clear temporal patterning.

The results of the period-sensitive keyword extraction provided in Figure 4 shows terms that are used significantly over-proportionately (positive values) and under-proportionately (negative values) to provide a more accurate depiction of changes in keyword use over time in the COOEE.

|  |
| --- |
| Figure 4: Keywords significantly over- and under-used in the COOEE by period. |

Figure 4 shows that terms related to the journey to Australia (e.g., *wales*, *wind*, *norfolk*) are used over-proportionately in the first period, terms related to establishing a home life (e.g.; *husband* and *children*) and differences between England and Australia (e.g., *chief*) occur in the mid-period while terms related to exploration are used over-proportionately in the later periods (e.g., *travelled*, *creek*, *water*, *horses*, *deg*). We now turn to the results of the topic modelling.

3.2 Topic Model

The initial unsupervised data-driven topic model detected six thematically coherent topics in 13,931 sentences – other number of topics were tested but did not provide the same degree of coherence in the topics. In a second step, the keywords provide by the unsupervised model were used as seed terms. Table 4 provides the ten terms that are most strongly associated with each topic according to the second seeded topic model.

Table 4: Top 10 keywords for each of the 7 topics detected by the LDA-based topic model (Topic 7 serves as a bin category).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Topic1 (family)** | **Topic2 (journey)** | **Topic3 (landscape)** | **Topic4 (exploration)** | **Topic5 (indigenous)** | **Topic6 (employment)** | **Topic7 (other)** |
| dear | travel | water | camp | nativ | good | great |
| home | wind | river | hors | black | work | will |
| mother | arriv | creek | camel | natives | week | everi |
| father | ship | hill | horses | chief | money | upon |
| family | board | australia | horse | aborigin | gold | time |
| son | ocean | mile | horseback | native | pay | without |
| will | land | miles | night | aborigines | will | present |
| letter | clock | north | start | they | make | person |
| know | south | countri | last | white | well | part |
| time | sail | found | return | like | year | make |

Table 4 confirms that the topic model arrived at six thematically coherent topics as can be seen by the key terms associated with each topic. Topic 7 (other) serves as a bn category for sentences that could not be classified properly. The change in the proportion of topics across time periods is shown in Figure 5.

|  |
| --- |
| Figure 5: Results of the LDA-based Topic Model – proportion of sentences associated with topics across periods. |

Figure 5 shows that the dominant topic in the earliest period (1788-1800) the journey-topic which is associated with terms like *travel*, *wind*, *arriv*, *ship*, *board*, *ocean*, etc. Interestingly, this topic is almost absent in later periods. In the second period (1801-1820), the family-topic gains prominence which is characterized by terms like *dear*, *home*, *mother*, *father*, *letter*, and terms relating to inner states such as *know*, *think*, and *hope*. In the later two periods (1861-1880 and 1881-1900), the landscape and exploration-topics notably gain dominance while other topics with the exception of the employment-topic loose in prominence. The proportion of sentences that could not be unambiguously assigned to a topic (see the proportion of Topic7\_other in Figure 5) decreases from early to late. Overall, we can see a shift from journey, to family and employment, to landscape and exploration. We now turn to the interpretation and discussion of the results in light of the research questions presented above.

4. Discussion

The current analysis of the *Corpus of Oz Early English* (COOEE) has unearthed intriguing and unexpected findings showing what the private writings of early Australian settlers revolved around. We will now evaluate the results in light of the research questions posed above.

Regarding RQ1, the analysis reports a remarkably high number of idiosyncratic keywords (2,560) which were used significantly more frequently compared to their use in letters written in Britain. Of these keywords, 22 had an effect size greater than a phi-value of .02 which indicated that they were indicative of concepts typical to the COOEE. The keywords are predominantly referring to concepts relating to the journey to Australia (e.g., *miles*, *water*, *wind*, *sydney*, *started*), exploration (e.g., *natives*, *camels*, *camp*, *found*, *men*), and landmarks (e.g., *river*, *creek*, *range*, *south*, *north*, *west*) (see Table 3). Figure 2 furthermore indicates that keywords referring to exploration such as *camp*, *miles*, *horses*, *creek*, river are more central than key terms relating to the journey to Australia which are captured by keywords such as *wind* or *ship* which indicates that the topic of exploration is more central and prototypical for the personal written discourse of early Australian settlers.

With respect to changes in the use of keywords over time (RQ2), Figure 4 shows that during the earlier two periods in the data (1788to 1800 and 1801 to 1820), terms relating to the journey from Great Britain to Australia (e.g., *wind*, *norfolk*, *wales*) are significantly overrepresented while terms relating to exploration (e.g., *mulga, camp*, *miles*) are significantly under-represented. In the mid-periods (1821 to 1840 and 1841 to 1860), phenomena typical of Australia and community building are over-represented (e.g., *husband*, *children*, *chief*) while terms relating to exploration remain under-represented (e.g., *camp*, *trees*, *creek*, *water*). In the later periods (1861 to 1880 and 1881 to 1900), terms and concepts relating to exploration (e.g., *horses*, *travelled*, *creek*, *water*, *noticed*, *deg*) are overused whereas terms relating to the journey to Australia such as *wold (world)*, *colony*, *dear*, and *ship* are underused. In short, we can detect a shift from a focus on the journey to Australia, over community building and getting to know the local circumstances, to exploration and conquering the land.

This shift is corroborated by the topic modelling which detected six thematically coherent topics and one topic which served as a bin category for sentences that could not be classified with confidence (RQ3). The six coherent topics are *family*, *journey*, *landscape*, *exploration*, *indigenous*, and *employment*. The seventh topic figures particularly prominent during the early periods suggesting that the topics people wrote about during these early years were less focused and coherent compared to writing during later periods. As to RQ4, the shift in topics mirrors the trend observed in the changing importance of keywords in that we can observe a majority of sentences dealing with the journey to Australia during the first two periods and a notable predominance of sentences referring to landmarks and exploration in the latter two periods (see Figure 5).

On a methodological note, the study and the results presented here exemplify how dialectology, historical linguistics, and studies describing the content of collections of electronic texts can profit from adopting computational methods developed in natural language processing to explore and summarize large amounts of textual data which cannot reasonably be south through using traditional close-reading techniques. In particular, the present study showed how keyword analysis can be applied to explore the lexical, semantic, and narrative content of text collections and how topic modelling can assist in detecting thematically coherent topics and themes in collections of private correspondence.

5. References

Blei, David M., Ng, Andrew Y. & Jordan, Michael I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993-1022.

Field, Andy, Miles, Jeremy & Field, Zoë. 2012. *Discovering Statistics Using R*. London: SAGE.

Lindquist, Hans. 2009. Corpus Linguistics and the Description of English. Edinburgh: Edinburgh University Press.

7. Corpora and Software

Benoit, Kenneth, Watanabe, Kohei, Wang, Haiyan, Nulty, Paul, Obeng, Adam, Müller Stefan & Matsuo, Akitaka. 2018. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software* 3(30): 774. doi: 10.21105/joss.00774.

Denison, David & van Bergen, Linda. 2003. *A Corpus of late 18c Prose (CoRD).* The University of Manchester.

Feinerer, Ingo & Hornik, Kurt. 2020. tm: Text Mining Package. R package version 0.7-8. https://CRAN.R-project.org/package=tm

Fritz, Clemens. 2007. *From Early English in Australia to Australian English 1788-1900*. Frankfurt: Peter Lang.

R Core Team. 2021. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Watanabe, Kohei & Xuan-Hieu, Phan. 2021. seededlda: Seeded-LDA for Topic Modeling. R package version 0.6.0. https://CRAN.R-project.org/package=seededlda

1. This number deviates slightly from the official word count as the definition or word and how words are counted here differ from the way words were defined in counted by the corpus compilers. We use decided to rely on our own counts as we believe them to be of higher accuracy. [↑](#footnote-ref-1)