**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 letter 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 The private writing section of the COOEE consists of 607 unpublished letters and diary entries written in Australia or by native Australians on travels, between 1788 and 1900.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 eight 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 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 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 letter section of the COOEE which consists of 405 unpublished letters 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 text analytic methods adopted from Natural Language Processing (NLP) can assist

To determine and extract distinctive keywords and topics, the present study uses quantitative

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.

2. Previous Research

The present study addresses these research gaps and adds to existing research by focusing on long-term change in the Irish English amplifier system and by specifically zeroing in on interdependencies between amplifiers and adjectives. In addition, the current study extends existing research by zeroing in on a well-circumscribed and restrictive variable context (similar approaches have been undertaken by D’Arcy 2015; Tagliamonte 2008; Tagliamonte and Roberts 2005; Tagliamonte and Denis 2014), namely, the variation of amplifier variants in pre-adjectival slots. Restricting the analysis to a narrowly defined variable context allows the precise evaluation of factors contributing to variation, which would not be possible if the analysis built on relative frequencies (see Labov 1966, 49; Tagliamonte 2011, 9-10).

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 2: What topics are written about in the COOEE?

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

3. 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 CoL18P, and describes the data processing while the second subsection describes the statistical methods that have been applied to the data.

3.1 Corpus Description and Data Processing

The entire COOEE comprises 1353 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 XXX)[[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 XXX).

|  |
| --- |
| 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 XXX 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 26,562 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.

3.2 Statistical Methods

The present study makes use of two types of analyses:

(1) keyword analysis both using 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 compared to the control;

(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.

3.2.1 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 corpora. As this resulted in a large number of tests, we used Bonferroni corrections to avoid an inflation of alpha-error (false positive results). 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.

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 The visualizations

3.2.2 Topic Modelling

The effect size measure reported here is the logged p-value (Stefanowitsch and Gries 2005). The values of this effect size measure inform about whether the amplifier and adjective repel or attract each other. Values below 0 indicate rejection while values above 0 indicate attraction and values around 0 do neither indicate preference nor rejection. In statistical terms, values below 0 show that an amplifier and an adjective occur less frequently together than would be expected by chance while values above 0 show that an amplifier and an adjective occur more often together than expected by chance. For the analysis, all adjectives other than dangerous, different, difficult, good, important, little, necessary, and new were collapsed into one category (other). The advantage of covarying collexeme analyses over similar methods to evaluate collocational attraction is that it is a very robust method as it is an extension of Fisher’s Exact Test that does not rely on distributional assumptions as tests form the χ2-family of tests do. The following section presents the results of the quantitative analysis.

4. Results

The final data consisted of 11,947 adjectives of which 13.98% were amplified (see Table 1). The most frequent adjective amplifier in the data is very with 1,230 instances which amounts to 73.65 percent of all instances of amplification.

4.1 Keyword Analysis

The keyword analysis detected 2,591 uni-, bi- or trigrams that occurred significantly more frequently in the COOEE compared to the CoRD.

Table 3: Overview of words significantly overused in COOEE (keywords) with phi > .02 (COOEE total = 312,281, Control total = 167,111)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Word** | **COOEE** | **Control** | **p-value (uncorrected)** | **x2** | **phi** | **Significance (Boferroni corr.)** |
| miles | 2,033 | 4 | 0.000 | 1,080.9 | 0.0475 | p<.001 |
| water | 1,505 | 87 | 0.000 | 606.5 | 0.0356 | p<.001 |
| country | 1,243 | 48 | 0.000 | 551.5 | 0.0339 | p<.001 |
| one | 2,475 | 499 | 0.000 | 430 | 0.0299 | p<.001 |
| us | 1,725 | 274 | 0.000 | 394.6 | 0.0287 | p<.001 |
| creek | 609 | 0 | 0.000 | 324.8 | 0.026 | p<.001 |
| river | 666 | 14 | 0.000 | 321.2 | 0.0259 | p<.001 |
| camp | 593 | 1 | 0.000 | 313.7 | 0.0256 | p<.001 |
| natives | 557 | 0 | 0.000 | 296.9 | 0.0249 | p<.001 |
| pm | 551 | 0 | 0.000 | 293.7 | 0.0248 | p<.001 |
| found | 918 | 102 | 0.000 | 277.1 | 0.024 | p<.001 |
| south | 517 | 2 | 0.000 | 270.4 | 0.0237 | p<.001 |
| north | 520 | 7 | 0.000 | 259.7 | 0.0233 | p<.001 |
| course | 627 | 33 | 0.000 | 258.2 | 0.0232 | p<.001 |
| like | 735 | 69 | 0.000 | 243.7 | 0.0225 | p<.001 |
| camels | 429 | 0 | 0.000 | 228.2 | 0.0218 | p<.001 |
| two | 1487 | 325 | 0.000 | 228.6 | 0.0218 | p<.001 |
| went | 716 | 79 | 0.000 | 216.7 | 0.0213 | p<.001 |
| wind | 421 | 5 | 0.000 | 211.6 | 0.021 | p<.001 |
| east | 407 | 3 | 0.000 | 209 | 0.0209 | p<.001 |
| west | 405 | 5 | 0.000 | 203 | 0.0206 | p<.001 |
| range | 377 | 0 | 0.000 | 200.4 | 0.0204 | p<.001 |
| men | 611 | 59 | 0.000 | 199.4 | 0.0204 | p<.001 |
| feet | 413 | 9 | 0.000 | 197.8 | 0.0203 | p<.001 |

Network below

|  |
| --- |
| Figure 4: Network co-occurrence graph of keywords in the COOEE. |

The second most frequent adjective amplifier is so with 283 instances representing 16.95 percent of overall amplification which goes to show not only that very is the dominant adjective amplifier in the HCIE data but that it is pronouncedly so as all other amplifiers combined make up merely 26.35% of all instances of amplification.

|  |
| --- |
| Figure 2: Relative frequency of selected keywords across periods in the COOEE. |

Figure 1 displays the rate of amplification among adjectives in the HCIE data and shows that the rate of amplification in attributive contexts is lower compared to the rate of amplification in predicative contexts and that rate of amplification in attributive contexts is highly stable while the rate of amplification in predicative contexts has steadily increased since 1750 after having experienced a drop in the earliest subsection of the data.

|  |
| --- |
| Figure 3: Keywords significantly over- and under-used in the COOEE by period. |

It should be borne in mind, however, that the percentage values depicted in Figure 1 are restricted to the variable context, i.e. not all adjective slots are considered. The dendrogram shown in Figure 2 displays the results of the Semantic Vector Space model. According to Figure 2, very and so are highly semantically similar based on their co-occurrence profiles with adjectives and stand apart from all other amplifiers. In addition, variants which are phonetically and orthographically similar (true and truly as well as real and really) are also reported to behave very similarly with respect to their collocational preferences.

4.1 Topic Model

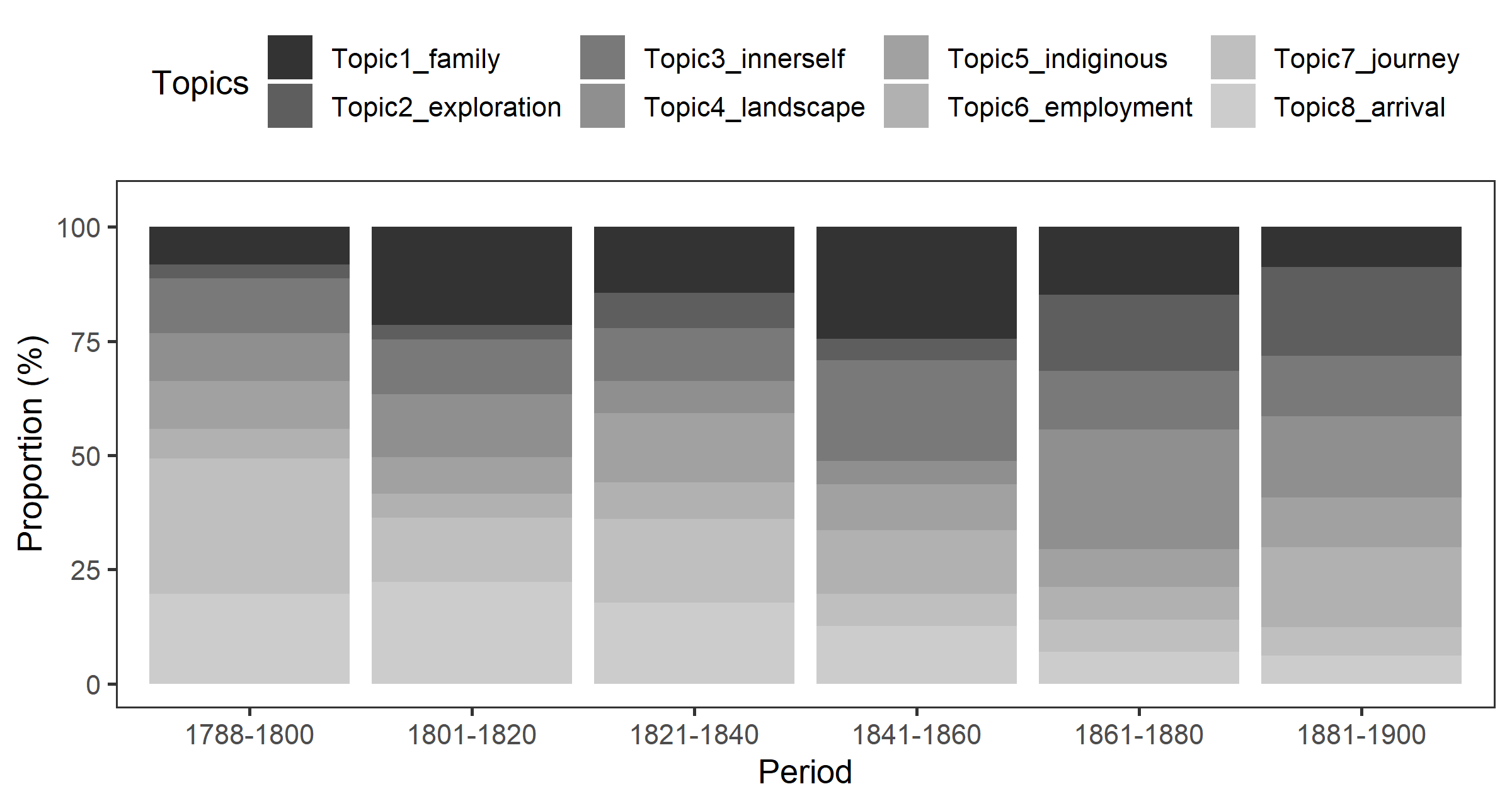
The topic model

Table 4: Top 10 keywords for each of the 8 topics detected by the LDA-based topic model.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Topic1 (family) | Topic2 (exploration) | Topic3 (innerself) | Topic4 (landscape) | Topic5 (indiginous) | Topic6 (employment) | Topic7 (journey) | Topic8 (arrival) |
| mr | water | think | miles | natives | good | day | great |
| dear | miles | can | country | made | get | night | country |
| mrs | camp | never | water | went | got | morning | upon |
| letter | horses | shall | river | saw | work | wind | time |
| well | found | say | north | like | money | rain | every |
| family | camels | well | creek | men | per | sunday | may |
| time | mr | see | course | came | year | ship | part |
| send | last | may | feet | white | go | weather | colony |
| mother | started | know | small | black | land | last | place |
| write | m | little | range | several | week | board | man |

According to Figure 3, the distribution of amplifier types in attributive contexts is remarkably stable. In predicative contexts, so has replaced very as the dominant adjective amplifier during the first decades of the 20th-century. In addition, so – which appears to be the sole rival of very in the Irish data – is almost categorically restricted to predicative contexts.

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



We now turn to the LD scores of adjective amplifiers to gain a better understanding of changes in the range of adjectives that variants co-occur with. The distribution of LD scores across real time is shown in Figure 4. In attributive contexts, the LD scores of very decrease over time while the LD scores of so and other amplifiers remain stable and only show some minor fluctuation. In predicative contexts, LD scores of very, so, and other adjective amplifiers decrease almost uniformly over time with very consistently having the lowest LD scores. Interestingly, there is no noticeable change in the patterning between the 1851 and 1930, i.e., during the time period when so has replaced very in predicative contexts.

5. Discussion

The current analysis of the *Corpus of Oz Early English* (COOEE) has unearthed intriguing and unexpected findings. The analysis of private letters of Australian emigrants shows a remarkably high number of idiosyncratic keywords. stable amplifier system in attributive contexts and the replacement of very by so as the dominant adjective amplifier in predicative contexts during the first decades of the 20th-century (see Figure 3).

One issue which remains unanswered in the present study relates to the role that social factors have played in the restructuring of the adjective amplifier system of Irish English. The inclusion of social factors into the analysis of adjective amplification is warranted because the "use of linguistic forms to increase distinctiveness of particular groups is a driving force for the acceleration of change" (Labov 2002, 19). Thus, future research should go beyond focusing on the interdependencies between amplifiers and adjectives and take a multivariate approach which simultaneously takes various intra- and extra-linguistic factors into account. In this respect, the current analysis represents a starting point for more in-depth studies on factors that cause so, rather than really, to become dominant in predicative contexts in the written domain.

6. References

Aijmer, Karin. "Are you totally spy? A new intensifier in present-day American English." Marqueurs discursifs et subjectivité. Ed. Sylvie Hancil. Rouen: Universités de Rouen and Havre, 2011. 155-172.

Blei DM, Ng AY and Jordan MI (2003) Latent dirichlet allocation. *Journal of Machine Learning Researc*h 3(3): 993–1022.

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

Xiao, Richard and Hongyn Tao. "A corpus-based sociolinguistic study of amplifiers in British English." Sociolinguistic Studies 1.2 (2007): 241-273.

7. Corpora and Software

Hornik, Kurt. Package "OpenNLP – Apache OpenNLP Tools Interface." 2016. Web. 10 October 2019.

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

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/.

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)