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

Martin Schweinberger (The University of Queensland, Artic University of Norway)

m.schweinberger@uq.edu.au

**Abstract**

This study takes a machine-learning approach to analyse the private letter section of the *Corpus of Oz Early English* (COOEE). While the COOEE consists of 1353 unpublished letters, published books, and historical texts written in Australia, New Zealand or Norfolk Island, or by native Australians on travels, between 1788 and 1900, this study focuses on the private letter section to analyse topics emigrants and early settlers of Australia talked about in their private correspondence. As such, the present study sets out to exemplifies how computational methods such as keyword extraction, topic modelling, and error analysis can assist historical linguists, dialectologists, and corpus linguists in unearthing patterns and topics that would be hard to identify using traditional methods. A Latent Dirichlet Allocation-based topic model reports that five topics figure prominent in the texts ranging from descriptions of new homes and living situations, over love letters, and military action, to family life, and contacts with indigenous people. 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

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, analyses of corpus data almost exclusively focus on the occurrence of specific features. This paper explores the content of a corpus, the *Corpus of Oz Early English* (COOEE) (Fritz 2004). The COOEE consists of 1353 unpublished letters, published books, and historical texts written in Australia, New Zealand or Norfolk Island, or by native Australians on travels, between 1788 and 1900. 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 *The English language of the north-west in the late Modern English period: A Corpus of late 18th century Prose* (Denison & van Bergen 2003).

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).

3. Data and Methodology

The following section consists of two subsections: the first subsection provides information about the COOEE corpus 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

While the 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, this study focuses on the private letter section to analyse topics emigrants and early settlers of Australia talked about in their private correspondence. The private letter section consists of 405 private letters amounting to 797,569 tokens (304,493 types) (see Table XXX).

Of these 405 letters, 27 were composed by women amounting to 6.7 percent. Letters were on average about 2,000 words long (1969.3). Due to the semi-literate status of the writers, orthography in the COOEE is variable and mirrors the writers’ or scribes’ pronunciation (see, for instance, the examples in (1) and (2

Table 1: Overview of number of letters and words per period in the COOEE.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Period** | **Author Gender** | **Letters (N)** | **Types (N)** | **Tokens (N)** |
| 1788-1800 | male | 9 | 18,033 | 63,568 |
| 1801-1820 | male | 46 | 33,332 | 102,608 |
| 1821-1840 | female | 1 | 170 | 295 |
| 1821-1840 | male | 86 | 48,063 | 114,444 |
| 1841-1860 | female | 6 | 6,858 | 19,934 |
| 1841-1860 | male | 131 | 88,065 | 213,955 |
| 1861-1880 | female | 8 | 6,576 | 17,965 |
| 1861-1880 | male | 33 | 29,876 | 74,056 |
| 1881-1900 | female | 12 | 20,562 | 59,286 |
| 1881-1900 | male | 73 | 52,958 | 131,458 |
| **Total** |  | **405** | **304,493** | **797,569** |

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

The style of the letters represents a hybrid of speech-like writing and formal phraseology. As such, the material is more similar to transcribed court proceedings which have been analyzed, for instance, by Claridge and Kytö (2014) compared to analyses of written texts proper. The relatively higher rate of poverty and, thus, semi-literate writers in and from Ireland – compared to authors of private letters from England, for instance – make the study of IrE particularly intriguing. This is so because texts produced by semi-literate writers are more speech-like than letters produced by authors who were taught to write in school. As such, semi-literate IrE resembles historical spoken language more closely than other sources of historical data (an exception being the Old Bailey Corpus which represents transcripts of spoken court proceedings, see again, for instance, Claridge and Kytö 2014).

As a final step, the data were manually cross-checked to minimize the amount of errors arising from the (semi-)automatic data processing.

3.2 Statistical Methods

The present study makes use of three types of analyses: (1) semantic vector space modelling to investigate semantic similarities of individual amplifier variants, (2) lexical diversity scores (LD) to investigate range of adjectives that amplifier variants co-occur with, and (3) covarying collexeme analysis to investigate changes in the collocational profiles of amplifiers. The latter two analyses are complementary in that lexical diversity scores provide information about the general profile of amplifier variants as it indicates their breath by showing how many different adjective types amplifier variants co-occur with while covarying collexeme analysis evaluate the collocational strength between individual amplifier variants and specific adjective types.

3.2.1 Semantic Vector Space Modelling

Semantic vector space models allow for the inspection of the relative similarity of amplifier types based on their co-occurrence frequencies with adjectives (also referred to as vectors or embeddings) (see Levshina 2015). The underlying reasoning is founded in distributional semantics, according to which, words that share many collocates are semantically similar (Stefanowitsch 2010, 368-370). In this view, semantic similarity can be measured in terms of the similarity of collocation profiles. The results of vector space models are commonly displayed in the form of dendrograms in which semantically similar elements inhabit the same branches.

3.2.2 Lexical Diversity

To analyse changes in the diversity of the collocational profiles of amplifiers in the HCIE, the current study uses a simple lexical diversity measure (LD). LD scores are calculated by dividing the number of adjective types a given amplifier co-occurs with by the number of tokens of that amplifier type (see 3).

(3) LD = NAdj.Types/NAmp. Tokens

LD can reach a maximum value of 1 in which case it indicates high lexical diversity. The lower the LD value, the lower the degree of lexical diversity (see Table 2).

Table 2: Example of LD calculation.

Amplifier Amp. Tokens (N) Adj. Types (N) Calculation LD value

variantA 10 1 1/10 .1

variantB 10 5 5/10 .5

variantC 10 10 10/10 1

very1675-1750 11 17 11/18 .647

very1751-1850 86 368 86/368 .234

very1851-1930 102 845 102/845 .121

As Table 2 shows, the LD values of very decrease over time because there are 11 tokens of very that co-occur with merely 17 adjective types in letters written between 1675 and 1750 while there are 102 tokens of very that co-occur with 845 distinct adjective types in letters written between 1851 and 1930. The LD value allows testing whether the lexical diversity of an amplifier variant changes over time and thus whether amplifier variants are semantically broadening, i.e. becoming more general, or narrowing, i.e. specializing on modifying fewer adjectives across real time.

3.2.3 Covarying Collexeme Analysis

To analyse changes in the attraction between the most common amplifier variants and selected individual adjectives, the present study makes use of covarying collexeme analysis which is part of the collostructional family of analyses (Gries and Stefanowitsch 2004; Hilpert 2006; Stefanowitsch and Gries 2003, 2005). Covarying collexeme analyses allow for the quantification and evaluation of attraction between elements that occur in two distinct slots within a specified construction. In the present case, the first slot represents the amplifier slot and the second slot represents the adjective slot. Each slot can be occupied by a variant from a set of potential candidates – the set of amplifier variants for the first slot and the set of adjectives for the second slot. Covarying collexeme analysis provides information about whether the likelihood of a certain variant in the first slot affects the likelihood of another variant from another set occuring in the second slot. In other words, it is more likely that nice occurs in the second slot given that really occurs in the first slot compared with another amplifier taking the first slot. The p-values reported by the covarying collexeme analysis are Bonferroni corrected to control the inflation of α-error rates. This is necessary because a huge number of tests were run which would have resulted in an inflation of false positive results (α-error) if the corrections had not been applied. The covarying collexeme analyses were performed for each period (1675 to 1750, 1751 to 1850, and 1851 to 1930). 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. 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.

Table 3: Overview of the cleaned and processed data.

Variant N % % (of Amp.)

Ø (not amplified) 10,277 86.02

very 1,230 10.30 73.65

so 283 2.37 16.95

pretty 62 0.52 3.71

real 17 0.14 1.02

truly 14 0.12 0.84

awefully 13 0.11 0.78

extremely 9 0.08 0.54

perfectly, really 5 (10) 0.08 (0.04) 0.6 (0.3)

exceedingly 4 0.03 0.24

dreadfully, remarkably, surely, terribly, unusually 2 (10) 0.2 (0.02) 0.12 (1.2)

certainly, dead, deeply, excessively, excellently, fully, greatly, imminently, miserably, perfect, sincerely, vastly 1 (12) 0.01 (0.12) 0.06 (0.72)

Total 11,947 100.0 100.0

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 1: Percent of amplified adjectives across syntactic contexts and time in the HCIE.

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.

Figure 2: Horizontal cluster dendrogram showing the semantic similarity of adjective amplifiers in the HCIE based on their co-occurrence frequencies with adjectives.

Figure 3 shows the distribution of very, so, and other amplifier types in attributive and predicative contexts in the HCIE across real time to enable a more fine-grained understanding of changes in amplifier use.

Figure 3: Distribution of amplifiers across syntactic contexts and time in the HCIE.

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.

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.

Figure 4: Changes in LD scores of very, so, and other adjective amplifiers across syntactic contexts and time in the HCIE.

There are at least two issues related to the current measurement and display of lexical diversity. Firstly, the LD scores do not control for frequency and, thus, the most frequent adjective amplifier is bound to have the lowest LD sore by chance alone. However, this also implies that the LD scores for so should dip when so experienced an increase in frequency which is not the case.

Secondly, the distributions shown in Figure 4 could be affected by changes in the token frequency of adjectives. In other words, the changes in amplifier use could be confounded by changes in the frequency of adjectives. Indeed, a variant such as so could become dominant, not because it undergoes some kind of change, but merely because the adjectives it collocates with become more frequent. To control for changes in adjective frequency as a confounding factor, the percentages of adjectives across real time are displayed in Figure 5.

Figure 5: Changes in the use of adjective types across syntactic contexts and time in the HCIE.

Figure 5 indicates that the adjective system in HCIE data is remarkably stable (adjective types other than the seven most frequent adjectives (dear, few, good, great, last, little, and old) were merged collapsed into the category other). The only trend that emerges from Figure 5 is a slow and steady increase in the use of infrequency adjectives. The stability of the adjectival system suggests that changes in the frequency of individual adjective types cannot be responsible for the consistency in the patterning of LD scores shown in Figure 4.

The covarying collexeme analysis did not detect any significant collocations after being Bonferroni-corrected, i.e. significance levels were adapted to control for repeated testing. This means that none of the combinations of amplifiers and adjectives occurred significantly more or less frequently than would be expected by chance. However, the analyses did reveal interesting changes in collocation strength (see Figure 6).

Figure 6: Collocation strength of adjective amplifiers by adjective type across syntactic contexts and time in the HCIE.

The panels in Figure 6 reveal that even though the amplifier system in attributive contexts is highly stable, there have been collocational changes below the apparently conservative surface. With respect to individual patterns, very is most strongly associated with good in both syntactic contexts and preferred by other adjectives in predicative contexts in the latest subsection of the data – as indicated by the trends in the solid lines. Other amplifiers strongly associate with bad and glad in predicative contexts and with other adjectives in attributive contexts during the period from 1751 to 1850. So has experienced an increase in collocational strength with glad and good in predicative contexts as well as with other adjectives in attributive contexts as shown by the upward trends in the dotted line. Given the results of the covarying collexeme analyses, the attraction between so and good in predicative contexts, although appearing to be only a minor increase in attraction, may quite likely be the reason why so has replaced very as the dominant amplifier in the most recent data due to good's sheer frequency. The fact that so increasingly associates with other adjectives in attributive contexts is also interesting as it poses the question of whether broadening in in one syntactic context could affect the occurrence of a variant in another context. A more thorough discussion of the findings is provided below.

5. Discussion

The current analysis of the historical development of adjective amplification in the HCIE has unearthed intriguing and unexpected findings. The analysis of Irish emigrant letters shows a remarkably 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). The stability of adjective amplification in attributive positions in itself appears to be a remarkable finding because it stands in stark contrast to previous claims. In fact, the stability of the adjective amplifier system in the HCIE does not conform with the notion that the domain of adjective amplification, or the domain of intensification more generally, is prone to change (Brinton and Arnovik 2006, 441; Ito and Tagliamonte 2003, 257; Quirk et al. 1985, 590), a site of invention and renewal (D’Arcy 2015, 450), and an area of grammar that undergoes "fevered invention" (Bolinger 1972, 18). Since Schweinberger (2020) found a very similar trend in historical fiction data in the COHA, the question arises if such statements can be maintained in their universality or whether the idea of constant renewal is constrained by extra-linguistic restrictions regarding genre or text type. This finding loses its surprising aspect, however, once we consider previous research that has focused on long-term diachronic change which showed that it is precisely predicative contexts that are the locus of innovation and the position of earlier stages of change (see de Smet 2012). As such, it is even expected that predicative contexts show change before innovations diffuse to the more conservative attributive contexts – not last because attributive contexts are more restrictive, for instance with respect to innovative variants such as so.

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.

Aijmer, Karin. "That’s well bad. Some new intensifiers in spoken in British English." Corpus Approaches to Contemporary British English. Eds. Vaclav Brezina, Robbie Love, and Karin Aijmer. New York and London: Routledge, 2018a. 60-95.

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.

R Development Core Team. "R: A language and environment for statistical computing." 2012. R Foundation for Statistical Computing. Vienna, Austria. Web. 10 October 2019.

Siemund, Peter and Lukas Pietsch. "The Hamburg Corpus of Irish English." unpublished. Universität Hamburg. Collaborative Research Center: Multilingualism. Project H5: Hiberno-English: Variation and Universals in Contact-Induced Language Change. Web. 10 October 2019.