A Computational Approach to Analysing Historical Irish Emigrant Letters

Martin Schweinberger

This study takes a machine-learning approach to analyse a corpus of Irish emigrant letters (The Hamburg Corpus of Irish English; HCIE). The HCIE is particularly suited for this study as the letters are written by semi-literates which has the advantage that spoken and non-standard features are more pervasive compared to more standardized and formal written data. The aim of the study is to exemplify how computational methods and natural language processing can assist corpus linguists in unearthing patterns in textual data that would be hard to identify using traditional corpus linguistic methods.

By using keyword extraction, topic modelling, and error analysis, this study shows which themes were prominent in the emigrant letters, it highlights changes in the prominence across time, and it allows drawing inferences about linguistic characteristics of the emigrants’ pronunciation as revealed by systematic orthographic errors in the writing of the semi-literate authors. For instance, the inclusion of sounds within words (epenthesis), monophthongisation, and dialectal features.

A Latent Dirichlet Allocation-based topic model reports that five topics figure prominent in the letters ranging from descriptions of new homes and living situations, over love letters, and military action, to family life, and contacts with indigenous people (Indians).

An additional case study on the frequency and use of non-standard second person personal pronouns (ye, youse, and yourself) finds remarkable use of non-standard forms with a notable decrease over time, indicating that the personal pronoun system has undergone a development towards standardization. While this development aligns with analogous historical changes in American English, it contrasts starkly with a cross-varietal trend in spoken registers where non-standard form of personal pronouns, remains one of the most prominent features across regional and non-standard varieties of English.

The results presented here show that so and very are semantically highly similar and that the increase in so is neither triggered by significant changes in its collocational profile nor by semantic broadening or an increase in its lexical diversity. The results are therefore interpreted to suggest that the increase in so will remain local rather than expanding to a wider variety of contexts and genres.