Angus McIntosh Centre for Historical Linguistics
THE UNIVERSITY of EDINBURGH
School of Philosophy, Psychology and Language Sciences

Corpus Linguistics in Scotland, Edinburgh 2.12.2016
'Diachrony through corpora'

Postgraduate session abstracts

The *Corpus of Sermons in Early Modern England*: Methodological Issues in Corpus-Based Idiolectal Analysis

Hiroshi Yadomi, University of Glasgow

I have compiled the *Corpus of Sermons in Early Modern England*, a single-genre corpus, consisting of the language of 20 preachers. The main focus of the corpus is investigation of idiolectal variation and its pattern in the Early Modern period. I propose that confessional states of sermon writers (Anglicans vs. Puritans) may explain the inter-personal variation of language use.

The presentation will include a brief description of the corpus as well as the motivation to compile it and the scope which the corpus can offer, followed by a few pilot studies to show that the corpus is useful to pinpoint inter-personal variation and its pattern.

The corpus has overcome various methodological challenges which idiolectal analysis entails: insufficient data and different sociolinguistic and pragmatic variables potentially affecting the language choice of speakers. The corpus contains c. 50,000 words of sermon texts written by a single preacher, whose sociolinguistic variables (gender, social class and age) are fixed as consistently as possible.

However, the corpus is still in development and carry some methodological problems. For example, sermons contain numerous Bible quotations which may skew the data. Additionally, we are not quite sure how precisely printed sermons reflect the language spoken from the pulpits of the same period. I will discuss these problems and grope for solutions.

Code-switching in late medieval council registers from Aberdeen

Anna D. Havinga, University of Aberdeen

The Aberdeen Council Registers – as Scotland's oldest and most complete run of civic records – are a rich source for the investigation of changes in legal concepts, practices and language. As part of a digital humanities project funded by the Leverhulme Trust, the first eight volumes of these registers (1398–1511) are being transcribed and annotated with the aim of creating a digital resource that is relevant to academics and the general public (see https://aberdeenregisters.org/). A particularly interesting issue to study for linguists is the occurrence of code-switching from Latin to Scots (and vice versa) and its diachronic development in these records. In this talk, I will examine this issue in the fifth volume of the Aberdeen Council Registers (1441–1468), a

volume which displays a significant amount of both intersentential and intrasentential codeswitching. The analysis of instances of language mixing will reveal when and how scribes switched between codes. I will also explore the reasons for code-switching in this corpus. Furthermore, the comparison of the frequency of code-switches during the almost 30 years covered by volume 5 will highlight the importance of diachronic corpora for linguistic research.

References to Authority in Early English Medical Writing
Karoliina Ollikainen, University of Glasgow

My PhD project investigates how references to authority changed in early English medical writing between 1375-1700 by combining quantitative corpus methods with qualitative analysis. During these 300 years English scientific discourse underwent a fundamental ideological shift which also affected the style of scientific writing. Middle English medical texts were built on the learned tradition of scholastic medicine, whereas the early modern period saw the gradual rise of empirical science which emphasized observation and experience. To account for the two different argumentation styles, scholastic and empirical, I am conducting a corpus search both on a set of communication verbs (e.g. SAY, WRITE) and mental verbs (THINK, BELIEVE) based on the observations made by previous studies. Each relevant result is then analysed within its context and categorised according to what the underlying ideology behind the reference is. My research examines both references to external authorities (e.g. "Hippocrates says X") as well as instances where the authors set themselves up as authorities ("I think that Y"). The material of my study comes from the Corpus of Early English Medical Writing compiled in the University of Helsinki.

Presenting Statistical Analyses to Historical Linguists
Daisy Smith, University of Edinburgh

The availability of large-scale diachronic corpora has made it possible to investigate historical linguistic phenomena using sophisticated statistical methodology. Gries (2015:97) describes corpus data in general as "observational and, thus, usually unbalanced and messy/noisy". He points out that techniques frequently employed in other areas of linguistics, particularly those which use experimental methods, can be fruitfully applied to corpus linguistics. More than that, he suggests that these methods are actively necessitated by the very nature of corpus data.

Historical corpus data is no exception to this generalisation, but Historical Linguistics is a field in which the scope for 'big data' analysis has been understandably smaller than in others. Compare, for example, a corpus of electronically-extracted social media interactions with a corpus of Middle English texts which must be painstakingly transcribed and digitized.

How, then, in a field where hand-drawn isoglosses are for many a not-so-distant memory, should a researcher embarking on a complex statistical analysis present his or her results to the historical linguistic community?

In this talk, I will present an analysis of data from *A Linguistic Atlas of Older Scots*, firstly considering what information can be gleaned from some basic descriptive statistics. Secondly, I will present the results of a Generalised Linear Model (GLM) fit to the data, showing how this can provide further insight into trends compared with basic statistical methods. Lastly, I will present a Generalised Additive Model (GAM) fit to the same dataset, again showing how this method improves insight into the data.

My aim in presenting at CLiS is to gain an understanding of how linguists working with diachronic corpus data, whether they have a statistical background or not, engage with the statistics I present. I hope that this understanding will enable me to maximise the accessibility and informativity of my statistical communication.

Reference

Gries, S. (2015). The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora*. 10(1). Edinburgh: Edinburgh University Press. 95-125.