Special Issue on Visualisations in Historical Linguistics: Introduction

Benjamin Molineaux, Bettelou Los, Martti Mäkinen

Special Issue on Visualisations in Historical Linguistics

Introduction

Benjamin Molineaux[1], Bettelou Los[1], Martti Mäkinen[2]

1 The University of Edinburgh, UK

2 Hanken School of Economics, Finland

Abstract
The advent of ever-larger and more diverse historical corpora for different historical periods and linguistic varieties has led to the impossibility of obtaining simple, direct —and yet balanced— representations of the core patterns in the data. In order to draw insights from heterogeneous and complex materials of this type, historical linguists have begun to reach for a growing number of data visualisation techniques, from the statistical, to the cartographical, the network-based and beyond. An exploration of the state of this art was the objective of a workshop at the 2018 *International Conference on English Historical Linguistics*, from whence most of the materials of this Special Issue are drawn. This brief introductory paper outlines the background and relevance of this line of methodological research and presents a summary of the individual papers that make up the collection.

keywords
data visualisation, historical linguistics, variation, change, corpus linguistics, historical dialectology

INTRODUCTION

This special issue is made up of a selection and expansion of papers from the *Workshop on Visualisations in Historical Linguistics*, held at the University of Edinburgh on 30th August 2018, as part of the *20th International Confrerence on Historical Linguistics* (ICEHL XX). The event was organised by Rhona Alcorn, Bettelou Los and Benjamin Molineaux of the Angus McIntosh Centre for Historical Linguistics (AMC), in lieu of the second AMC Symposium, which would have otherwise been held in the same year and institution as ICEHL XX, a sub-optimal overlap.

The Workshop's aim was to showcase a variety of recently developed tools to aid data visualisation in historical linguistics. Nine papers were presented at the Workshop, covering topics ranging from the statistical and theoretical underpinnings of visualisation, to the creation and application of diverse tools to specific historical linguistic problems. Some of the papers presented novel tools, others demonstrated the application of existing tools. Six of the presenters have submitted revised and expanded versions of their work to the current volume. The seventh paper (Hessle and Kirk) was recruited from another Edinburgh-based event: *Scots Words and Phrases in the Contemporary World: Back to the Future*, held on the 8th and 9th April, 2019. This invited paper presents visualisations of lexicographical and cartographical approaches to *The Linguistic Atlas of Scotland* data that make use of geographical mapping, a technique familiar to the field in the form of historical atlases (like, for instance, the Atlas of Late Middle English (*LALME*, began 1952), [McIntosh et al. 1986]), but transformed by the dynamic applications of advances in technology. Its inclusion makes for a well-rounded, internally-cohesive set of materials addressing the achievements and possibilities of this burgeoning area of studies.

The motivation for the specific focus of our special issue and the workshop it developed out of is the proliferation of larger corpora and data sets for the various historical periods, locations, genres and registers of languages (here the focus is on English and Scots), which provide us with much better coverage, but make individual variables increasingly difficult to situate, explore and understand. As a result, summarising, filtering and organising data in ways that facilitate more accurate and nuanced generalisations of our data are particularly welcome.

Visualisations can provide, richer, more intuitive perspectives on the relations between linguistic and extra-linguistic variables (temporal, spatial and social distributions), as well as on relations between different levels of linguistic structure (sounds-spellings, syntax-semantics, etc.). The papers in this special issue of the JDMDH discuss give a technical overview of the visualisations tools they have developed or re-purposed, and provide a concise application of these to a particular problem in the history of English and Scots.

In an important historical linguistic visualisations paper, [Martin Hilpert 2011] remarks that "As yet, there is no established catalogue of analytical techniques that are both adapted towards the special characteristics of diachronic corpus data and that are designed to address the types of theoretical questions that matter to linguists interested in diachronic change." The papers in this collection, we believe, show that important advances are underway in this area, and that, while no full catalogue of such visual analytical techniques has been established in the interim years, we are certainly on a road to refining the key theoretical and methodological issues that shall underpin them.

I CONTRIBUTIONS

The opening contribution to this special issue is Herman **Moisl**'s paper on statistical techniques for visualising high-dimensional data. As its subtitle implies, the article constitutes *a roadmap* for linguists exploring materials that are not easily reducible to standard two-dimensional graphic representations. Indeed, the size of —and variation within— historical linguistic corpora means that a large variety of linguistic variables must be reduced to a manageable number in order to provide meaningful insights and feed into hypothesis generation. After reviewing some key data concepts, two such techniques are explored and evaluated: dimensionality reduction and cluster analysis. A bespoke corpus of key English-language historical texts is compiled and explored via frequency of letter co-occurrences. Through the above techniques, different measures of textual proximity are calculated for the corpus, showing a close match to our previous knowledge about the historical and linguistic provenance for the texts. Overall, the paper provides unexperienced researchers with a very readable introduction to the statistical background of high-dimensional data visualisation, alongside a crucial sanity-check for researchers already applying these methods in their work.

Like Moisl's paper, the second contribution to our collection also takes a step back from current practices in historical data visualisation, to question a number of assumptions and propose methodological innovations. **Hessle and Kirk** focus on the traditional dialectological and lexicographic material from the *Linguistic Atlas of Scotland* (LSA) [Mather and Speitel, 1975; 1977], calling into question the criteria for mapping the presence or absence of features within a cartographic space where variation is inherent. The problem of data representation in survey data is considered such as relates to the idea of responses representing hyponomy vs. synonymy or the distinction between 'standard' and 'dialect' forms and their binary opposition as presence/absence. The paper goes on to propose a novel perspective on the data using the mathematical notion of the ‘excluded middle’. Further observations and recommendations for data digitisation and colour-representation of the mapped patterns are put forward and exemplified through the University of Vienna’s digitisation of the LSA project.

In the third paper of this collection, **Beck and Butt** showcase a Visual Analytics system, *HistoBankVis*. This is a methodological tool which can aid investigations into linguistic, and particularly syntactic change by providing visualisations of correlations between linguistic features in data sets. *HistoBankVis* allows investigators to zoom out by providing at-a-glance overviews, but also allows zooming in to individual data points and their underlying corpus annotations. The accuracy of the*HistoBankVis* tool depends on the quality and consistency of the annotation methods of the individual corpora employed; even the Penn-style treebank annotation system, which has become somewhat of an industry standard, is not wholly consistent from one corpus to the next. The paper discusses specific issues with respect to data uncertainty and annotation standards which need to be solved in order to make the most optimal use of *HistoBankVis* – and demonstrate how *HistoBankVis* can itself be an important tool to help overcome these problems.

Lubbers and Los chart the development of style in a single genre from the 16th century to the present day by using existing tools on a corpus of selections from horse manuals on the same topic: how to feed your horse. As the aim is to focus on the syntax rather than on the lexicon, the selections are tagged with part-of-speech (POS) labels and then stripped of lexical content, so that only a sequence of POS-tags and punctuation marks remain. N-grams of these sequences are subjected to various correspondence analyses and to a hierarchical clustering analysis. The resulting visualisations correctly identify the relative chronological order of the texts purely on the basis of the differences between them, as well as the n-gram sequences that drive this ordering. As these differences cannot be due to changes in the grammar – the syntax of English has by and large remained unchanged since the 16th century – , they can be taken as indicators of how the stylistic conventions of this genre develop.

Martti **Mäkinen**’s exploratory article repurposes a tried and tested stylometric method and script – *Stylo* for R – in the study of Middle English documents. The paper adopts the middle ground between the relational linguistic space applied in [LALME] and the real linguistic continuum of medieval England by imposing an unattended analysis of character 3-grams on texts drawn from the corpus of *Middle English Local Documents* ([MELD]) compiled at the University of Stavanger. The paper discusses the usability of character n-grams in identifying diatopically conditioned “fingerprints” of texts and grouping them according to the similarities in the distribution of the n-grams over the corpus texts. The results show that 3-grams and *Stylo* can be a usable and effective tool in grouping texts automatically according to their regional varieties. The novel use of the stylometric tool may help to open doors for other computational approaches to historical dialectology.

In their contribution, **Schlüter and Vetter** present a powerful new web-based tool for the exploration of raw Google Books Ngrams data, christened *(an:a)*-lyzer. Purpose-built with the *Shiny* R package, the interface allows for a flexible and speedy management of a truly enormous dataset encompassing over 468 billion words across five centuries, allowing users to categorise lexemes based on custom phonological and etymological factors. The application is currently tailored to a key problem in the historical phonology of English: the restoration of consonantal corelates of word-initial <h>, as diagnosed by the allomorphy of a preceding indefinite article (*a/an*). The paper guides readers through the possibilities of the app and the numerous visualisation and statistical summaries that it makes available. The possibilities of visually tracking categories, subcategories and individual lexical items across such a vast dataset allows users to find patterns in the data with fine-grained internal structure conducive to lexical diffusion and exemplar-based studies, amongst others. The presentation of the resource as both a standalone, downloadable tool and as a web-based application, furthermore, provides maximal access and flexibility, hopefully boosting research on additional aspects of the available dataset. Most importantly, however, the tool serves as a proof of concept for applications that can manipulate and visualise extremely large 'messy' data sets for hypothesis generation and testing.

Our issue concludes with a paper that applies visual mapping of sound-spelling relationships to a concrete problem in the historical development of Scots. **Molineaux, Maguire, Karaiskos, Alcorn, Kopaczyk and Los** present a grapho-phonological mapping tool —*Medusa*— as applied to the *From Inglis to Scots* (Alcorn *et al., forthcoming*) corpus data, compiled, in turn, from the texts and metadata of the *Linguistic Atlas of Older Scots* [Williamson, 2014]. The case study compares the reconstructed sound-substitution sets for the complex graphemic representations <ch; cht; th; tht> mapping their lexical distributions and taking readers through evidence from etymology, phonological typology, palaeography and historical orthography. The result is a novel reconstruction of the underlying sound values for each one of the target items in the record, alongside a series of sound and spelling changes that account for the data.

Acknowledgements

As editors we would like to thank all the authors included in the current issue for their continued support and enthusiasm towards bringing together these materials in the midst of the various challenges and delays imposed by the COVID 19 pandemic. Our thanks also go out to the thirteen external reviewers that greatly improved the quality of the individual contributions. We are also grateful to the main editor of the JDMDH, Laurent Romary, for agreeing to host this unique set of papers, and facilitating and encouraging us through the inevitable technical challenges of editing an online, open-access journal.

References

FITS = Alcorn, R., Honkapohja, A., Karaiskos, V., Kopaczyk, J., Los, B., Maguire, W. & Molineaux, B. (compilers). *From Inglis to Scots: A Corpus of Grapho-phonological Correspondences (1380-1500) with Associated Corpus of Changes*. The University of Edinburgh (Edinburgh). Forthcoming. http://www.amc.lel.ed.ac.uk/fits/

Hilpert, M. Dynamic visualizations of language change: Motion charts on the basis of bivariate and multivariate data from diachronic corpora. *International Journal of Corpus Linguistics* 2011; 16/4: 435–46.

LALME = McIntosh, A., Samuels M. L., & Benskin, M. (compilers). *A Linguistic Atlas of Late Mediaeval English*. 4 vols. Aberdeen University Press (Aberdeen), 1986.

LAOS = Williamson, K. (compiler). *A Linguistic Atlas of Older Scots, Phase 1: 1380–1500*. The University of Edinburgh (Edinburgh), 2008. http://www.lel.ed.ac.uk/ihd/laos1/laos1.html

Mather, J.Y and Speitel, H.-H. (eds.). *The Linguistic Atlas of Scotland. Scots Section. Volume 1*. Croom Helm (London), 1975.

Mather, J.Y and Speitel, H.-H. (eds.). *The Linguistic Atlas of Scotland. Scots Section. Volume 2*. Croom Helm (London), 1977.

MELD = *The Middle English Local Documents Corpus*, version 2017.1. University of Stavanger (Stavanger), 2017. https://www.uis.no/research/history-languages-and-literature/the-mest-programme/a-corpus-of-middle-english-local-documents-meld/.