Universiteit Leiden

Language of the Poor in Late Modern Scotland: From the Archive to the Internet

Applicants

Supervisor Name	**Discipline**
Dr. Jelena Prokic	Computational linguistics, Leiden University Centre for Digital Humanities
Dr. Mo Gordon	Historical sociolinguistics, Leiden University Centre for Linguistics

Project description

Historical sociolinguistic researchers need access to as many different types of sociolinguistic data as possible to study language use in past societies, e.g. texts produced by different social layers of society, in formal and informal styles, as well as authors from different regions. Though texts produced by higher social classes are readily available in archives and online, the biggest challenge is to find material that allows us to investigate the language of the lower layers of society at times when access to education and literacy was socially stratified. Once retrieved from the archives, e.g. in the form of pauper petitions, an additional challenge is to digitize these data into a format that allows researchers to carry out (socio)linguistic analyses. Due to the restricted schooling and literacy of the lower orders, the handwriting is often difficult to decipher and the spelling idiosyncratic. This poses challenges to the accuracy of automatic spelling standardizers and OCR tools, while at the same time the relatively small number of historical documents at our disposal does not allow for the application of tools like Transkribus or machine learning techniques in general.

Within this context, the project aims to digitize Late Modern Scottish pauper petition letters —usually written by members of the lower orders of society — and to investigate what digital tools are most suitable for this purpose.

More precisely, the aim of the project is two-fold:

1. To make the language of lower-class writers from Late Modern Scotland (1700-1900) digitally available and suitable for historical (socio)linguistic research. Scottish English of this period is of particular interest in light of the prestige variety that emerged in the South of England and that contributed to the further Anglicization of the Scottish writing system.[1] To date, relatively little is known about how Anglicization affected the lower orders.
2. To explore what digital tools and methods can be used to digitize and annotate these data to make them useful for various types of historical (socio)linguistic research.

[1] See for instance: Devitt, A. J. (1989). Standardizing Written English: Diffusion in the Case of Scotland. Cambridge University Press.
Dossena, Marina (2005). Scotticisms in grammar and vocabulary: ‘Like runes upon a standin’ stane?’. John Donald.
Meurman-Solin, Anneli (1997). Differentiation and Standardisation in Early Scots, in Jones (ed.), The Edinburgh history of the Scots language (pp. 3–23). Edinburgh University Press.

Research Trainee Profile

The main aim of the traineeship will be to gain insights into the challenges and solutions in digitizing historical sociolinguistic data. For example, the Text Encoding Initiative (TEI) provides guidelines for standardized machine-readable texts, but each kind of text and research requires its own specific subset of TEI annotations and mark-up. As part of this, a specific corpus manual will have to be compiled for the corpus of pauper petitions. Trainees are also challenged to think about what information historical texts can provide and how this can be translated into metadata and annotations. For instance, if they are interested in finding out about the use of hypercorrect prestige forms, they can devise a way to annotate occurrences of these forms in texts so that they can be retrieved electronically. For this purpose, they will be introduced to tools for digital corpus annotation and hone their palaeographical skills. As the corpus will be designed for historical (socio)linguistic data, familiarity with historical (socio)linguistic research methods are a must. Trainees will also be given the opportunity to formulate their own research questions relating to the Scottish pauper petition material and to think of a set of annotations that will facilitate their research.

Since trainees need to have an awareness of the methodological and annotation-related implications of historical (socio)linguistic research questions, our preference goes out to ResMA students who generally have more experience with empirical research and the formulation of research questions. However, third-year BA students or MA students are also invited to apply if they demonstrate to have the required skills/experience.

Collaboration

Over the past decades, digital information technologies have opened up many new research avenues and have triggered new research questions; they have also allowed us to reconsider findings that were based on much smaller data sets, simply because research was largely carried out manually. One of our aims is to bring together digital technologies and historical (socio)linguistic research to facilitate interdisciplinary research within the humanities. The cross-pollination lies in the intersection between computational methods and more traditional archival work that is typically associated with historical data. Furthermore, experience with digital information processing increasingly becomes a requirement for jobs within the humanities and outside, so via this project we hope to familiarize students with digital tools, as well as the challenges and prospects they offer when working with historical data. Students will also gain an awareness of how data are processed electronically, which will not only be of use in the field of historical (socio)linguistics but for many types of research within the humanities and beyond, as well as for the digital representation of historical documents. The combined expertise in text digitization, as well as the experience with the development of databases (Jelena Prokic) and the experience with digital text representation and corpus annotation for historical sociolinguistic research (Mo Gordon), will help students develop digital skills that facilitate their own research goals, as well as those of the research community at large.

Deliverables

The intended end result of this project is that the trainees will have digitized a set of texts that will be made available to the research community at large:

a. As a starting point, the applicants have already got access to c. 24 photocopies (c.65 pages of text) of Scottish pauper petitions that are ready to be digitized. An additional number of circa 100 petitions have already been localized in the archives, and more will follow in the future.

b. An important first step of designing the corpus and providing relevant metadata is to consider what kind of questions the corpus should address. Each trainee will therefore have to write a brief research proposal on a topic relating to the anglicization of written Scottish English and the corpus material in question.

c. Together with their supervisors, the trainees will have devised an annotation system that will facilitate historical (socio)linguistic research, but they should also consider what annotations could benefit the larger research community, including neighbouring disciplines. The resulting annotated corpus should be accompanied by a corpus manual that will allow other researchers to navigate and search the corpus as well.

d. The manual, including a corpus description, and the corpus data (if possible) will be made available online via the University's web host.

Planning

February-March: rendering transcriptions into digital format, trainees familiarize themselves with the data
March-April: reading relevant background material, proposing research questions based on data and background reading
April-May: digital tool training stage (annotation, mark-up), making decisions about what metadata are relevant
May-June: annotate digital transcriptions (based on proposed research questions)
June-July: writing up a description of the annotations and the corpus in form of a corpus manual
July-August: making materials available online

There will be at least one meeting a month to discuss progress with the whole team, and brief weekly meetings with one of the supervisors to discuss practical matters.

Student Application

The application should include a short CV and a motivation letter in which you elaborate on your experience with historical (socio)linguistic research, e.g. relevant papers that you may have written, relevant courses taken, and methodologies that you have applied so far, general research interests, etc. Affinity with digital (corpus) tools is much appreciated, as are familiarity with Scottish English, written standardization and/or prescriptivism.

Interested students should send their applications to Mo Gordon (m.s.gordon@hum.leidenuniv.nl) and Jelena Prokic (j.prokic@hum.leidenuniv.nl)