Workshop: Statistics in historical corpus linguistics

4th-5th October 2019

Room 1.37, Iontas Building, North Campus

Maynooth University, Maynooth, Ireland

This workshop is hosted by the *Chronologicon Hibernicum* Project (ERC Horizon 2020, grant no. 647351), and is funded by the National University of Ireland Pilot Early Career Academic Scheme. We have invited seven speakers to talk about **the application of statistical methods to language corpora, especially to those of historical languages** (please see the titles and abstracts below).

This workshop is open to the public and you are more than welcome to attend. Registration is free and includes tea/coffee during the breaks and casual lunches on the 4th and the 5th. However, attendees other than the invited speakers are advised to arrange their own travel and accommodations, for which an information sheet is attached. Please register in advance via the following link:

https://forms.gle/TzwXWjfkQAoLURTU6

Please do not hesitate to contact the organisers if you have any questions:

Fangzhe Qiu (fangzhe.qiu@mu.ie)

Ellen Ganly (ellen.ganly.2013@mumail.ie)

Programme:

4th October 2019

9:30-9:45	*Opening Remarks*
9:45-10:45	Marco Aquino-López and David Stifter (Maynooth University): "Statistical methods in the Old Irish language: A methods point of view of the ChronHib project"
10:45-11:05	*Tea / Coffee Break*
11:05-12:05	Robin Ryder (Ceremade, Paris Dauphine): "Phylogenetic Models of Language Change: Validating Interference and Quantifying Uncertainty"
12:05-13:30	*Lunch (at the venue)*
13:30-14:30	Martin Hilpert (Univesité de Neuchâtel): "Variability-Based Neighbor Clustering with Historical Corpus Data: Results, New Applications and Future Directions"
14:30-14:50	*Tea / Coffee Break*
14:50-15:50	Anthony Kroch (University of Pennsylvania): "Recent Results in Quantitative Diachronic Syntax"
15:50-16:50	Ann Taylor (University of York): "The Independence of

	Information Status and Syntactic Change: The Case of OV to VO in the History of English and Icelandic"
19:00 - Dinner at 'The Avenue', Main Street, Maynooth	

5th October 2019

10:00-11:00	Søren Wichmann (Leiden University): "Things to do with a Hictorical Lexical Frequency Corpus"
11:00-12:00	George Walkden (University of Konstanz): "Detecting Syntactic Change and Stability"
12:00-13:00	*Closing remarks and lunch*

Abstracts:

Statistical methods in the Old Irish language: A methods point of view of the ChronHib project.

Marco Aquino-López and David Stifter (Maynooth University)

This talk will discuss the primary goals and challenges of the project 'Chronologicon Hibernicum - A Probabilistic Chronological Framework for Dating Early Irish Language Developments and Literature' as observed from a statistical point of view. A discussion on the database created by the project and its structure will lead to an analysis of the potential statistical methods that can be used to help us understand the transition from the Old Irish language into Middle Irish.

In this talk, we will show some results obtained using Bayesian Statistics on particular case studies (Old Irish *inna/na* and *etar/iter*). Bayesian logistic regression with measurement errors was applied to these cases, which allowed us to observe the probability of these variations over the period between 750 and 870 A.D. This will display the advantages of using these techniques over traditional methods.

Phylogenetic models of language change: validating inference and quantifying uncertainty

Robin Ryder (CEREMADE - Paris Dauphine)

Since Gray & Atkinson (2003), phylogenetic models have become a common tool to reconstruct the history of language diversification over the past few millennia, especially when using cognate data. Usually, these models assume the history of some related languages can be represented by a tree, and inference procedures are built to infer that tree as well as the age of proto-languages, often in a Bayesian framework.

I will discuss the statistical challenges specific to such studies: how do we know that the data are useful and trustworthy? How do we choose a model, and ensure that our modelling assumptions do not introduce systematic bias? How do we quantify our uncertainty, and decide which inferences are solid enough? Most of the examples in the talk will be based on Sino-Tibetan languages, whose prehistory remains controversial, with ongoing debate about when and where they originated; I will discuss the methodology we employed (Sagart et al., 2019) to establish cognates, infer relationships between languages and estimate the age of their origin.

Variability-based neighbor clustering with historical corpus data: Results, new applications and future directions

Martin Hilpert (Université de Neuchâtel)

Variability-based neighbor clustering (VNC) has been devised as a method to automatically identify stages of language change in temporally ordered corpus data (Gries & Hilpert 2008, 2012). Its main advantages are that it offers a data-driven, inductive way of periodizing historical corpora and that it lets the analyst decide on the linguistic phenomenon that forms the basis for a given periodization. This talk will take stock of some of the work that has been done with VNC. It will be explained how the method works, what results have been obtained, and what research questions are still left to be explored.

The most common use of VNC takes historical frequency trends in order to partition diachronic corpus data into stages. One promising application of VNC that deviates from this has targeted the analysis of diachronic changes in morphological and syntactic productivity. In Hilpert (2013), changes in the usage of formations with the nominalizing suffix -ment have been used as a basis for a comparison between historical stages in the development of that suffix. Perek and Hilpert (2017) have further extended this line of research by partitioning the history of a grammatical construction according to qualitative stages of productivity. In a study of the "Verb the hell out of NP"-construction, it is shown that the semantic development of a construction does not always match that of its quantitative aspects, like token or type frequency. In another study, historical data illustrating the way-construction affords a comparison between a VNC-based assessment of productivity changes with results of a collostructional analysis.

With regard to future directions of VNC, this talk will take up other corpus-based measures, as for example dispersion, which up to now have not been used as a basis for periodization, but which reflect characteristics of linguistic forms that are highly relevant for the analysis of linguistic change.

References:

Gries, Stefan Th. & Martin Hilpert. 2008. The identification of stages in diachronic data: variability-based neighbor clustering. Corpora 3/1, 59-81.

Gries, Stefan Th. & Martin Hilpert. 2012. Variability-based neighbor clustering: a bottom-up approach to periodization in historical linguistics. In Terttu Nevalainen and Elizabeth Closs Traugott (eds.), The Oxford handbook of the history of English. Oxford: Oxford University Press, 134-144.

Hilpert, Martin. 2013. Constructional Change in English: Developments in Allomorphy, Word Formation, and Syntax. [Studies in English Language] Cambridge: Cambridge University Press.

Perek, Florent & Martin Hilpert. 2017. A distributional semantic approach to the periodization of change in the productivity of constructions. International Journal of Corpus Linguistics 22/4, 490-520.

Recent Results in Quantitative Diachronic Syntax

Anthony Kroch (University of Pennsylvania)

Due to the rapid evolution of the technology for Natural Language Processing, we linguists are now at the frontier of a new era in corpus linguistics. We have, or soon will have, access to data sets that are larger by several orders of magnitude than what have heretofore been available. In this talk, I will present some recent studies by students and collaborators at the University of Pennsylvania that show the promise of such "big data" for the study of the interface between grammar and usage, at both the diachronic and synchronic levels. The cases that I present will show the power of simple mathematical analysis to reveal strikingly regular patterns in the grammatical and lexical choices that speakers/writers make in the course of language production. Some of these patterns are cross-

linguistically valid and can perhaps be applied to the analysis of corpora of a broad range of languages and even in situations where the size of the corpora is limited.

The independence of information status and syntactic change: the case of OV to VO in the history of English and Icelandic

Ann Taylor (University of York)

The aim of this talk is to question the idea that the change from OV to VO in English is caused by or related to information structure in any way, as proposed by, for example, Hróarsdóttir (2009). The analysis is based on data from seven Old English and three Early Middle English texts extracted from two syntactically annotated corpora. The results show that while the syntax of the OE/EME VP is changing over time, the effects of IS remain constant. In a model with two post-verbal object structures, a conservative one where VO order is associated with new information and an innovative one where VO has no particular IS constraints attached, we find that, over time, as the proportion of the innovative structure increases, the proportion of new objects in post-verbal position decreases until it approaches the proportion of new information objects in the text as a whole, the result of the high proportion of new objects in post-verbal position in the conservative structure being increasingly diluted by the lack of an IS effect in the innovative structure. Reanalysed data from Hróarsdóttir 2009 shows that the quantitative patterns our model predicts show up even more clearly in Icelandic, a related language undergoing the same change from OV to VO.

Hróarsdóttir, Þorbjörg. 2009. OV languages: expressions of cues. In Hinterhölzl, Roland and Svetlana Petrova (eds.), *Information Structure and Language Change: New Approaches to Word Order Variation in Germanic*, 67-90. Berlin, New York: Mouton de Gruyter.

Things to do with a historical lexical frequency corpus

Søren Wichmann (Leiden University)

Using illustrations from analyses of the Google N-Gram and COHA lexical frequency data I will show how diachronic information on lexical frequencies can be used to gain new insights into the dynamics of both meanings and shapes of words. For instance, frequency data can be used to register societal upheavals and to establish links between shifts in attitudes and lexical-semantic change tendencies, and they also lend themselves to uncover formal trends in the creation of new lexical items. Along with the presentation of results such as these, I will also discuss methodologies issues, including the use of word embeddings (cooccurrence vectors) and a particular algorithm for detecting periods of establishment and obsolescence of words.

Detecting syntactic change and stability

George Walkden (University of Konstanz)

The claim that parataxis precedes hypotaxis in language is often found in the literature on language change. One version of this claim is the notion that subordinate clauses become proportionally more frequent over time. In this talk I present evidence from several annotated historical reference corpora (English, Icelandic, French, Portuguese, Irish, Chinese) suggesting that this claim is false. The focus will be on the methodological and statistical aspects: what types of model are appropriate for evaluating claims like this, and how do we evaluate the models themselves?