Maynooth Spring Symposium on Computational and Corpus-based Linguistics
Hosted by the Chronologicon Hibernicum Project

April 4, 2017
John Hume Board Room (3rd Floor)

9:30 - 10:15: Marieke Meelen (University of Cambridge)

Towards a Welsh historical treebank

In this talk I will discuss some of the challenges I faced annotating this morphologically rich historical corpus, some specific to Welsh (or other Celtic languages), others inherent to working with historical corpora in general. Apart from challenges, there are also promising opportunities for future research. I will argue for a standardized extension of the tag sets currently used for the Historical English and Icelandic corpora, for annotating languages like Welsh with very rich inflection in the form of verbal or prepositional suffixes. I will explain the workflow from preprocessing texts to POS tagging and chunk parsing and shed light on how the present annotated corpus can easily be queried and furthermore extended, not only with further texts, but also with additional features in the future so that it is useful for both philologists and linguists.

10:15 - 11:00: Christopher Yocum

Text Clustering and Methods in the Book of Leinster

Abstract: This paper will use methods developed within the discipline of machine learning and statistical analysis to accomplish two goals: first, to demonstrate the means and methods of unsupervised machine learning techniques in early Irish literature and second, to discuss the implications of the application of this methodology to the Book of Leinster with a view towards a larger research project.

11:00 - 11:15
Coffee Break

11:15 - 12:00: Aaron Ecay

Statistical Approaches to Old English Syntax

12:00 - 14:00
Lunch Break

14:00 - 14:45: Theodorus Fransen (Trinity College Dublin)

Towards a computational lexical resource for the diachronic study of Irish verbs

The research described in this paper employs a computational approach to investigate

Tionscnamh "Chronologicon Hibernicum" faoi chomairce ERC, Roinn na Sean-Ghaeilge,
Ollscoil Mhá Nuad, Maigh Nuad, Co. Chill Dara, Éire.
ERC-Project "Chronologicon Hibernicum", Department of Early Irish,
Maynooth University, Maynooth, Co. Kildare, Ireland.

E chronhib@nuim.ie **W** maynoothuniversity.ie/chronhib

erc
European Research Council
Established by the European Commission

historical developments in the Irish verbal system, which is subject to major changes between Early Irish (7th-12th centuries) and (Early) Modern Irish (post-12th centuries). At the same time, there is insufficient digital support to systematically trace these developments. An invaluable resource is the electronic Dictionary of the Irish Language (eDIL), covering the language from ca. 700 until ca. 1700. However, this dictionary is based mainly on Early Irish sources. Moreover, no comprehensive (digital) lexical resource currently exists for the modern period (post-1700). Consequently, we are faced with problems of both coverage and continuity in terms of lexicographical support for the historical periods of Irish.

Fortunately, work is currently underway to compile a dictionary for the modern period, i.e. the Royal Irish Academy's Foclóir na Nua-Ghaeilge 'Dictionary of Modern Irish' project, covering the period 1600-2000. The project employs state-of-the art computational methods including a Part-of-Speech tagger for modern standard Irish, which is in the process of being adapted to facilitate recognition of pre-standard forms in the corpus. This method allows one to search for a modern lemma and find all the inflected forms and associated historical variants in the corpus.

This paper will report on further adaptation of the modern-language tagger by the author to facilitate recognition of earlier verb forms, while at the same time investigating possibilities for creating tagging tools for Old Irish (7th-9th centuries) to project forward to modern forms. This approach includes the making available of interconnected, linguistically annotated corpora, thereby remediating the above-mentioned issues with diachronic lexicographical coverage and continuity. It is hoped that the research output will better facilitate and indeed accelerate the diachronic study of Irish verbs.

14:45 - 15:30: Oksana Dereza (National Research University, Moscow)

Lemmatizing Old Irish: Lexicons, Rules and Neural Networks

The task of lemmatization, being one of the basic steps in Natural Language Processing, is still far from being solved for underresoursed languages with rich morphology and inconsistent orthography such as Old Irish. I will give an outline of several methods I used to create an Old Irish lemmatizer and demonstrate its performance on different types of historical Irish data.

15:30 - 15:45
Coffee Break

15:45 - 16:30: David Willis (University of Cambridge)

Geospatial variation and diffusion in the history of Welsh

16:30 - 17:45: Discussion Period