TNA ACTIVITY REPORT

PROJECT TITLE:

From Modern to Early Irish: Retrogressive Diachronic Morphological Tagging Methods and UD Tagset Interoperability Solutions based on Detailed Linguistic Analysis

Author: Dr Theodorus Fransen

Current position: Postdoctoral Researcher

Affiliation: Data Science Institute, University of Galway, Ireland

Host institution: Institute of Formal and Applied Linguistics, Charles University, Czech Republic

Mentor(s): Dr Silvie Cinková

Period of stay: 21 September – 21 December 2022

Introduction

The stay financed by the CLS infra TNA scheme was used to work together with Dr Silvie Cinková at ÚFAL, Charles University Prague, in order to explore morphological tagging and syntactic parsing of texts according to the Universal Dependencies (UD) scheme, going from the Modern Irish period (c. 18th c.–present) back to Early Irish (7th–12th c.). Due to various issues and complexities found with corpus annotation for Modern Irish, and the philological challenges relating to Classical Modern Irish poems, it was decided to change the initial objective and restrict the annotation to about 1000 words of Early Modern Irish (c. 13th–17th c.). The fellowship has resulted in a publicly accessible, morphologically and syntactically labelled syllabic poem from c. 1655 entitled *Mo mhallacht ort, a shaoil* (edited and translated by Mac Cárthaigh 2013), which is in the process of being merged with a UD corpus of pre-standard Irish texts (Scannell 2022), currently adhering to UD v2.11 (https://github.com/UniversalDependencies/UD_Irish-Cadhan).

Methodological plan

The main contribution described in the initial proposal was the creation of a treebank of about ~5000 tokens from Old and Middle Irish (= Early) as well as Early Modern Irish texts, employing the pipeline described in Scannell (2022), adapting the Irish UD POS tagset and features.

Upon the start of the fellowship, the mentor and fellow decided upon the following plan:

1. Get familiar with the UD annotation guidelines and the ones written specifically for Irish
2. Liaise with Prof Scannell in relation to best practice for obtaining a standardised and parsed version of the corpus that the TNA fellow was to work on
3. Choose and get familiar with a CoNLL-U editor

CLS INFRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984

4. Check a “first-pass” standardised version of the texts, and manually correct wrongly modernised forms
5. Use a projecting parser (Scannell 2022) to obtain a (noisily) annotated first version of the treebank, i.e. project modern tags and dependency relations back to the original text on the basis of a corrected standardised version of the text
6. Manually correct the projected morphological tags and dependency relations on the basis of the guidelines for Irish, using the chosen editor — and publish the corpus

Description of the research visit and its outcomes

It was especially point 4. and 6. above that made both the mentor and fellow realise that the task of morphologically and syntactically annotating 5000 tokens was too big a task for a three-month fellowship. It was decided to restrict the work to one time period (Early Modern Irish) and one text (a Classical Modern Irish or ‘bardic’ poem), roughly containing 1000 tokens. It did not prove critical to adapt the morphological tagset for the text chosen, but interesting limitations and solutions were identified should earlier Early Modern Irish texts be added to the corpus — that is, when the anchor point remains Modern Irish, with mapping historical variants to modern lemmas and using e.g. the simplified case system of the contemporary language. Further details in relation to the annotation challenges and choices made are documented in a dedicated GitHub repository reflecting the fellow’s work, available at https://github.com/ThFransen84/UD_Irish-Bardic. This GitHub repository contains an important issue (https://github.com/ThFransen84/UD_Irish-Bardic/issues) in relation to annotation conventions that need to be either rectified in the existing Modern Irish treebanks, or better documented, and the solution will have repercussions for the fellow’s Bardic treebank as well. The fellow’s treebank is in the process of being merged with Scannell’s Cadhan Aonair pre-standard Irish corpus, available at https://github.com/UniversalDependencies/UD_Irish-Cadhan.

The fellow has gained invaluable knowledge and skills during his fellowship, including:

- The background and motivation behind the Universal Dependencies scheme and the main people behind its development, as well as related activities and personnel in ÚFAL.
- Morphological and syntactic annotation practice according to the UD scheme, and specific guidelines/conventions for Irish
- Gaining familiarity with the structure of the CoNLL-U format for tab-separated token-based indexing and annotation of sentences, and the popular ConlluEditor (Heinecke 2019) to query and graphically manipulate CoNLL-U files
- Using UDpipe (Straka 2018), an interface to which is available through LINDAT/CLARIAH-CZ.
- A better understanding of Early Modern Irish syllabic (bardic) poetry, the linguistic features and conventions of this genre, and Irish syntax in general (admittedly taking up much more time than originally envisaged)

Considerations for future work

As addressed in previous sections, dealing with multiple Irish time periods and texts, in combination with challenges of tagset interoperability, provided too much work for one person in three months. However, challenges in annotating older texts using modern-language labeling conventions were instrumental for the fellow’s understanding of potential future pitfalls in aligning morphological tagging schemes from a diachronic perspective. The fellow has identified multiple issues/bugs in the

modern treebank for Irish, which are in the process of being reported back to the relevant treebank maintainers.

Evaluation of NLP tools used

The fellow was lucky enough to be able to rely on earlier work by Scannell, who built an orthographic standardiser for pre-standard Irish texts (Scannell 2014) and devised a projecting parser (Scannell 2022). The orthographic standardiser, which is based on rule-based as well as statistical methods, converts a pre-standard Irish word to its contemporary orthographic equivalent. This output is input to a projecting parser which uses pre-standard/standard word alignments and a parser trained on the contemporary Irish UD treebank to project the obtained tags for the standardised text back to the original text. However, the resulting CoNLL-U files currently still need a lot of manual correction, and it is here that the linguistic expertise of the fellow proved invaluable. The (noisy) output was corrected using the annotation software ConlluEditor (Heinecke 2019). UDpipe (Straka 2018) was used to compare various constructions in Irish syntax to get a better idea of the annotation conventions for Irish. Seeing that the fellow needed to interact with the existing treebank by Scannell (2022) published on GitHub, he greatly improved his skills using the distributed version control system *git*. He also became familiar with raising (treebank annotation) issues, which constitute an important outcome of this fellowship, and which are publicly visible on the Cadhan Aonair GitHub page.

References

Johannes Heinecke. 2019. ConlluEditor: a fully graphical editor for Universal dependencies treebank files. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 87–93, Paris, France. Association for Computational Linguistics.

Eoin Mac Cárthaigh. 2013. *Mo mhallacht ort, a shaoghail* (c. 1655): dán is a sheachadadh. In *Ériu* 63, pages 41–77.

Kevin Scannell. 2014. Statistical models for text normalization and machine translation. In *Proceedings of the First Celtic Language Technology Workshop*, pages 33–40, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Kevin Scannell. 2022. Diachronic Parsing of Pre-Standard Irish. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 7–13, Marseille, France. European Language Resources Association.

Milan Straka. 2018. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.