Developing an inflectional lexicon for Old Irish

Theodorus Fransen[1], Cormac Anderson[2], and Sacha Beniamine[3]

[1]Insight SFI Research Centre for Data Analytics, Data Science Institute, National University of Ireland, Galway
[2]Max Planck Institute for Evolutionary Anthropology, Leipzig
[3]University of Surrey, Guildford

KEYWORDS: Old Irish, inflectional lexicon, grapheme-to-phoneme conversion, computational morphology, finite-state transducers

This paper describes an inflectional lexicon of Old Irish nouns, and the tools developed for its creation. While Old Irish (c. 600–900 A.D.) is extensively documented, it remains digitally under-resourced. We develop a morphological description in the form of a fully inflected lexicon of Old Irish nouns, provided in both phonemic and orthographic notation. This entailed devising a computer-assisted, systematic, and reproducible grapheme-to-phoneme conversion pipeline and generating morphological forms through a finite-state transducer. We report on the considerable challenges posed by Old Irish in terms of its morphophonological complexities and its intransparent and inconsistent orthography. The inflected lexicon we develop will better enable computational studies in Old Irish morphology, further research into diachronic developments, and have a wide range of Natural Language Processing (NLP) applications.

Despite the fact that "Old Irish is the earliest period of Irish — or of any Celtic language — for which the extant record is sufficiently full and varied to permit a full synchronic description" (Stifter, 2009, p. 59), the language still lacks the range of digital resources available for other Indo-European languages (e.g., Latin, see Pellegrini and Passarotti, 2018). While there are a number of independent projects focusing on Old Irish lexicography (Griffith, Stifter, and Toner, 2018), the most comprehensive resource, both in terms of contemporary source material included and the level of grammatical annotation, is *Corpus PalaeoHibernicum* (CorPH) 'Old Irish Corpus' (Stifter et al., 2021). However, in spite of the richness of the linguistic annotation in CorPH, it cannot be used as the basis for a morphological generator without considerable preprocessing, due to its inconsistent orthography for lemmata and the way it segments complex morphological structures.

Old Irish presents many challenges for the development of computational resources. The language has a complex phonology, an elaborate system of morphophonological alternations, and intricate patterns of morphological inflection (Anderson, 2016; Stifter, 2009; Thurneysen, 1946; Pedersen, 1909–1913). Further to this, the orthography is neither transparent nor consistent, and considerable differences in orthographic practice exist (Ó Cróinín, 2001). This complicates the development of a tool for automatic orthography-to-phonology conversion, as many orthographic sequences can have multiple readings; for instance, combinations of sonorant and stop are ambiguous, in that <rc> can represent /rg/ or /rk/ and <rg> /rg/ or /rɣ/, which we resolve by a) a normalised orthography, and b) some manual pre-processing.

We developed a pipeline for the creation of an inflectional lexicon. We began by extracting noun lemmata from the Old Irish Würzburg glosses (Kavanagh, 2001) and then devised a set of rules for orthography-to-phonology conversion, subsequently implemented using the Python package `Epitran` (Mortensen, Dalmia, and Littell, 2018). The resulting transcriptions act as the data input for a finite-state transducer (FST) adapted from Fransen (2019), allowing us to generate inflected forms of Old Irish nouns. Finally, we derived orthographic forms (and their variants) by applying conversion rules in the opposite direction. While this study focused on the Old Irish nouns in the Würzburg glosses, we intend to extend the lexicon by applying this pipeline to further corpora and other parts-of-speech.

This inflected lexicon makes possible systematic studies in data-driven morphology and typology (Pellegrini, 2020; Beniamine, Bonami, and Luís, 2021; Beniamine, 2021). It will also facilitate future research into diachronic and diatopic variation in Irish and the development of further NLP applications for the language. Moreover, the FST created to generate inflected forms provides a concise and thorough grammatical description of the Old Irish noun, and the automatic phonemic transcription rules can easily be re-used.

References

Anderson, Cormac (2016). “Consonant colour and vocalism in the history of Irish”. PhD thesis. Uniwersytet im. Adama Mickiewicza w Poznaniu. URL: https://hdl.handle.net/10593/14780.

Beniamine, Sacha (2021). “One lexeme, many classes: inflection class systems as lattices”. In: *One-to-Many Relations*. Ed. by Berthold Crysmann and Manfred Sailer. Berlin: Language Science Press.

Beniamine, Sacha, Olivier Bonami, and Ana R. Luís (2021). “The fine implicative structure of European Portuguese conjugation”. In: *Isogloss* 7.9, pp. 1–35. DOI: https://doi.org/10.5565/rev/isogloss.109.

Fransen, Theodorus (2019). “Past, present and future: Computational approaches to mapping historical Irish cognate verb forms”. PhD thesis. Trinity College Dublin, The University of Dublin. URL: https://github.com/ThFransen84/OIfst.

Griffith, Aaron, David Stifter, and Gregory Toner (2018). “Early Irish Lexicography – A Research Survey”. In: *Kratylos* 63.1, pp. 1–28. DOI: https://doi.org/10.29091/kratylos/2018/1/1.

Kavanagh, Séamus (2001). *A Lexicon of the Old Irish Glosses in the Würzburg Manuscript of the Epistles of St. Paul*. Ed. by Dagmar S. Wodtko. Mitteilungen der Prähistorischen Kommission 45. + 1 CD-ROM. Wien: Verlag der Österreichischen Akademie der Wissenschaften. DOI: 10.1553/0x0001fb6e.

Mortensen, David R., Siddharth Dalmia, and Patrick Littell (May 2018). “Epitran: Precision G2P for Many Languages”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari (Conference chair) et al. Miyazaki, Japan: European Language Resources Association (ELRA).

Ó Cróinín, Dáibhí (2001). “The earliest Old Irish glosses”. In: *Mittelalterliche volkssprachige Glossen. Internationale Fachkonferenz des Zentrums für Mittelalterstudien der Otto-Friedrich-Universität Bamberg 2. bis 4. August 1999*. Ed. by Rolf Bergmann, Elvira Glaser, and Claudine Moulin-Fankhänel. Germanistische Bibliothek 13. Heidelberg: Winter, pp. 7–31.

Pedersen, Holger (1909–1913). *Vergleichende Grammatik der keltischen Sprachen*. 2 Vols. Göttingen: Vandenhoeck & Ruprecht.

Pellegrini, Matteo (2020). “Using LatInfLexi for an Entropy-Based Assessment of Predictability in Latin Inflection”. English. In: *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*. Marseille, France: European Language Resources Association (ELRA), pp. 37–46. URL: https://aclanthology.org/2020.lt4hala-1.6.

Pellegrini, Matteo and Marco Passarotti (2018). “LatInfLexi: an Inflected Lexicon of Latin Verbs”. In: *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)* (Turin, Italy, Dec. 10, 2018). Ed. by Elena Cabrio, Alessandro Mazzei, and Fabio Tamburini. Vol. 2253. CEUR Workshop Proceedings. Aachen. URL: http://ceur-ws.org/Vol-2253/paper23.pdf.

Stifter, David (2009). “Early Irish”. In: *The Celtic Languages*. Ed. by Martin Ball and Nicole Müller. Hoboken: Routledge.

Stifter, David et al. (2021). *Corpus PalaeoHibernicum (CorPH) v1.0*. URL: http://chronhib.maynoothuniversity.ie.

Thurneysen, Rudolf (1946). *A Grammar of Old Irish*. Trans. by Daniel A. Binchy and Osborn Bergin. Revised and enlarged edition. Dublin: Dublin Institute for Advanced Studies. Repr. 1993, with supplement.