Alignement des approches de lemmatisation pour construire une base de connaissances en Données Linguistiques Liées et Ouvertes pour le vieil irlandais

Theodorus Fransen[1] Federico Simone Samperi[2] Elisa Roma[2] Paolo Ruffolo[1]
Marco Passarotti[1]
(1) CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Largo Gemelli 1, 20123 Milan, Italy
(2) Università di Pavia, Corso Strada Nuova, 65, 27100 Pavia, Italy
theodorus.fransen@unicatt.it,
federicosimone.samperi01@universitadipavia.it,
elisa.roma@unipv.it, paolo.ruffolo@unicatt.it,
marco.passarotti@unicatt.it

RÉSUMÉ

La présente contribution donnera un aperçu du projet MOLOR, qui vise à relier entre eux les textes, les lexiques et les données de flexion du vieil irlandais (env. 600-900 de notre ère) en utilisant les principes des Données Linguistiques Liées et Ouvertes (Linguistic Linked Open Data — LLOD). Cet article examinera les efforts numériques passés et présents pertinents pour soutenir la langue numériquement et se concentrera sur la collecte et l'harmonisation des formes canoniques comme première étape d'une architecture interconnectée de ressources linguistiques. La base de connaissances (BC) qui en résultera devrait bénéficier à un large éventail de chercheurs intéressés par la période médiévale irlandaise et au-delà.

ABSTRACT

Aligning lemmatisation approaches to construct a Linguistic Linked Open Data knowledge base for Old Irish

The current contribution will give an overview of the MOLOR project, which aims to interlink texts, lexicons, and inflectional data for Old Irish (c. 600–900CE) using Linguistic Linked Open Data (LLOD) principles. This paper will discuss relevant past and present digital efforts to support the language digitally and will focus on the collection and harmonisation of canonical forms as a first step in an interlinked architecture of linguistic resources. The resulting knowledge base (KB) is expected to benefit a wide range of scholars interested in the medieval Irish period and beyond.

MOTS-CLÉS : vieil irlandais; Données Linguistiques Liées Ouvertes; lemmatisation; morphologie.

KEYWORDS: Old Irish; Linguistic Linked Open Data; lemmatisation; morphology.

1 Background

While there has been a number of projects focusing on early medieval Irish lexicography (Griffith *et al.*, 2018), few have aspired to work towards harmonising entries in lexical resources and integrating textual and lexical resources. This absence can at least partly be explained by the complexities of Old Irish inflection, including intricate morphophonemic alternations and apophony (Stifter, 2009), complicated by an opaque orthography, tokenisation[1] and segmentation challenges — spacing in manuscripts often reflects the phonological and not the syntactic word — as well as lack of attestation. This presents challenges for lexicographic resources and natural language processing (NLP) — not least due to grammatical and spelling variation found in headwords both within and across dictionaries. To give an example, existing lexical resources (discussed below) provide different headwords for the originally deponent Old Irish lexeme 'to praise': pres. ind. 3sg. *molathair*, *molaithir*, *molaid* and *molaid(ir)*; none of these actually constitute the attested form in Old Irish, which instead is (active) *molid*, a spelling variant of (unattested) *molaid*.[2]

The *Electronic Lexicon of Medieval Irish* was developed as part of a PhD thesis by Nyhan (2006) as a response to limitations in the hard-copy *Dictionary of the Irish Language* or DIL (Marstrander *et al.*, 1913 1976), meanwhile digitised and revised as *electronic Dictionary of the Irish Language* or eDIL (Toner *et al.*, 2019), covering the language c. 700–1700CE. Nyhan's thesis aims to restructure a digitised subset of this dictionary using XML to enhance search capabilities, particularly for identifying inflected medieval forms with precision. Unlike the original DIL, where inflected forms were scattered, Nyhan's Lexicon organises these forms hierarchically. Nyhan also explored interlinking this Lexicon with the *Corpus of Electronic Texts* (CELT) (Ó Corráin *et al.*, 1997), allowing users to look up words in the Lexicon while reading texts in CELT. Unfortunately, both the Lexicon and the prototype for this linking remain unavailable due to broken hyperlinks.

The goal of the *Linking Dictionaries and Texts* (LDT) project (2003–2007),[3] a collaboration between the University of Ulster and University College Cork, was to enable interoperability between eDIL and CELT. The project included the creation of automated links between bibliographical citations in eDIL and corresponding texts in CELT (Nyhan, 2006). This would allow researchers to access texts and their contexts directly from words in eDIL. The *Digital Dinneen* project (2005–2008)[4] aimed to further support research on the evolution of the Irish language (Nyhan, 2008). It involved a digitised, XML-encoded edition of *Foclóir Gaedhilge agus Béarla* (Irish-English Dictionary) (Dinneen, 1927), allowing users to trace modern Irish words back to their historical forms in eDIL and the Electronic Lexicon of Medieval Irish, facilitated by an index of medieval Irish and Modern Irish headword

[1] For machine-learning based tokenisation efforts cf. Doyle *et al.* (2019).

[2] A Wiktionary discussion on this very headword reflects issues when choosing a suitable canonical form in Old Irish: https://en.wiktionary.org/wiki/Talk:molaidir

[3] https://celt.ucc.ie/LDT.html

[4] https://celt.ucc.ie/digineen.html

mappings (de Bhaldraithe, 1981).[5] Again, hyperlinks are broken and both the LDT and Digital Dinneen project similarly never fully seem to have come to fruition.

Independent but somewhat similar efforts have been carried out by Caoimhín Ó Donnaíle, who has worked on integrating dictionaries and texts utilising three tools for enhancing the use of online dictionaries: Multidict, Wordlink, and Clilstore, all available at `multidict.net`. Funded by European Commission projects, these tools aim to improve language learning, particularly for minority languages such as the Celtic languages (Ó Donnaíle, 2014). Using some of this linking technology, Ó Donnaíle has made available a list of 5000 conjugated Old Irish verb forms (King *et al.*, 2006) and maintains a comprehensive and freely-available network database of cognate words called *Bunadas*,[6] containing 14771 Old Irish entries.[7]

The most comprehensive lexical resource for Old Irish is *Corpus PalaeoHibernicum* or CorPH (Stifter *et al.*, 2021), one of the major outputs of the ERC-funded *Chronologicon Hibernicum* project (2015–2021). The aim of this project was "to develop a statistical methodology of linguistic dating in order to more precisely date the diachronic development of the Early Irish language" (Lash *et al.*, 2020, 1–3). While CorPH contains 10505 lemmas from 77 deeply annotated texts, the way it segments complex morphological structures means that linking back to the source text would be an almost impossible task.

Goidelex (Anderson *et al.*, 2024) is a novel open-access lexical database for Old Irish. Structured as a relational database in `csv` format, it currently contains 671 headwords derived from the important 8th-century Würzburg glosses (Kavanagh & Wodtko, 2001), which are not in CorPH. Goidelex contains fine-grained morphological and phonological annotations, including normalised orthographic representations and automated phonemic transcriptions. It is interoperable with existing Old Irish lexical resources like CorPH and eDIL and is fully compatible with the Paralex and CLDF standards (Beniamine *et al.*, 2023; Forkel *et al.*, 2018). For the purposes of the present paper it is important to note that Paralex also provides an RDF OntoLex-Lemon (McCrae *et al.*, 2017) conversion script.

2 Current work

Despite the above-mentioned digital efforts, the available resources for medieval Irish remain largely unconnected, leading to a fragmented lexicographic landscape. The MOLOR project described in this paper seeks to harmonise Old Irish canonical forms and make linguistic resources interoperable using controlled vocabularies and standards in electronic lexicography. Methodologically, the project draws heavily upon best practices obtained in the ERC-funded

[5] Meanwhile digitised and made searchable by Kevin Scannell at `https://cadhan.com/droichead/index-en.html`

[6] `https://www3.smo.uhi.ac.uk/teanga/bunadas/`

[7] `https://pure.uhi.ac.uk/en/projects/bunadas`

LiLa project: Linking Latin (2018–2023),[8] which focused on integrating distributed lexical and textual resources and NLP tools for Latin through the Linked Data framework. This approach utilises shared ontologies, data categories, communication protocols, and data models based on the Resource Description Framework (RDF) (Cyganiak *et al.*, 2014) to facilitate federated queries across diverse resources, in line with what Tim Berners-Lee has termed the Semantic Web (Berners-Lee *et al.*, 2001). By employing the Linked Data paradigm, the project aims naturally align with the FAIR principles of data management (Wilkinson *et al.*, 2016).

The current contribution will focus on and outline the design challenges and decisions involved in creating a collection of canonical forms or Lemma Bank for Old Irish, similar to the one developed in LiLa for Latin. This will be done by integrating canonical forms from both legacy and novel resources adopting the OntoLex-Lemon model (McCrae *et al.*, 2017), the *de facto* standard for modelling lexicographic data in RDF as well as by utilising extensions made to OntoLex-Lemon in the LiLa project. Serving as a lexical hub in a graph-based architecture of interlinked lexical resources and texts for Latin, a Lemma Bank was found to be indispensable in the LiLa project (Passarotti *et al.*, 2020).

The first version of the MOLOR Old Irish Lemma Bank, expected at the end of 2024, is projected to include some 5000 noun lemmas extracted from CorPH (Stifter *et al.*, 2021) and Goidelex (Anderson *et al.*, 2024). Goidelex is also being envisaged as the primary high-resolution morpho(phono)logical resource linked to the MOLOR Lemma Bank (Fransen *et al.*, 2024). Since the focus in Goidelex has so far been on nouns, the extraction of verb headwords and their inflectional features was instead entirely based on CorPH, supplemented with headwords from the Würzburg glosses dictionary (Kavanagh & Wodtko, 2001). Headwords were manually aligned, yielding 1100+ lemmas, estimated to cover the majority of verbal forms found in contemporary Old Irish manuscripts. Additional parts of speech will be incorporated in the near future, with the adjective possibly being next in line. Challenges in the compilation of such a collection of canonical forms include balancing linguistic granularity and representativeness due to Old Irish inflectional and spelling variation. The oral presentation will exemplify some complex Old Irish grammatical features and current design choices and solutions, i.e. how linguistic dimensions are mapped onto the linked data modeling.

3 Conclusions and future work

This paper has discussed the fragmented and unconnected nature of lexicographic resources for Old Irish (c. 600–900CE) and the collection and harmonisation of canonical forms — resulting in a Lemma Bank similar to the one developed in the LiLa project (Passarotti *et al.*, 2020) — as a first step in creating a LLOD KB for this language period. This endeavour

[8] https://lila-erc.eu/

is not without challenges due to linguistic variation inherent in Old Irish, in turn affecting LLOD modelling choices.

Upon completion, this Lemma Bank is expected to facilitate the seamless movement between texts and linguistic resources: querying federated resources using SPARQL (Prud'hommeaux & Seaborne, 2008) to — in the words of Tim Berners-Lee — "discover new things" (Berners-Lee, 2006). Linking Goidelex to the Lemma Bank, for example, would enable, through federated querying, the retrieval of fine-grained inflectional data from corpora to better understand Old Irish morphology. Other, perhaps more distant possibilities may include creating a bridge between two recently established Universal Dependencies[9] treebanks for Old Irish, referred to in Doyle & McCrae (2024), and valency information in specialised lexical resources such as PaVeDA (Luraghi *et al.*, 2024). As such, the MOLOR project will greatly assist scholars in elucidating the intricacies of Old Irish grammar and facilitate and speed up both time-intense philological work and computational tasks.

Looking ahead even further, a LLOD knowledge base for Old Irish could function as a solid foundation for creating an interlinked architecture for the medieval Irish period as a whole,[10] which in turn could facilitate the inclusion of the modern Gaelic languages derived from Old Irish — Modern Irish (Gaelic), Scottish Gaelic and Manx. Beyond (historical) Irish, the MOLOR project aims to serve as a model for lexico-morphological LLOD solutions for other low-resourced (historical or modern) languages with linguistic features such as complex morphology and a variable orthography.

References

ANDERSON C., BENIAMINE S. & FRANSEN T. (2024). Goidelex: a lexical resource for Old Irish. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, p. 1–10, Torino, Italia: ELRA and ICCL.

BENIAMINE S., ANDERSON C., CARROLL M., GUZMÁN NARANJO M., HERCE B., PELLEGRINI M., ROUND E., SIMS-WILLIAMS H. & TRESOLDI T. (2023). Paralex: a dear standard for rich lexicons of inflected forms. In *International Symposium of Morphology*. https://www.paralex-standard.org.

BERNERS-LEE T. (2006). Linked Data. https://www.w3.org/DesignIssues/LinkedData. Accessed: 2024-09-21.

BERNERS-LEE T., HENDLER J. & LASSILA O. (2001). The Semantic Web. *Scientific American*, **284**(5), 34–43.

[9] https://universaldependencies.org/

[10] Indeed, according to Stifter (2009, 59), Old Irish "furnishes a yardstick with which to assess the abundant literary production of the medieval period".

CYGANIAK R., WOOD D. & LANTHALER M. (2014). *RDF 1.1 Concepts and Abstract Syntax*. W3C recommendation, W3C. https://www.w3.org/TR/rdf11-concepts/. Accessed: 2024-09-21.

DE BHALDRAITHE T. (1981). *Innéacs Nua-Ghaeilge don Dictionary of the Irish Language*. Volume 1 de Deascán foclóireachta. Baile Átha Cliath: Acadamh Ríoga na hÉireann.

DINNEEN P. S. (1927). *Foclóir Gaeilge agus Béarla*. Dublin: Irish Texts Society, 2nd édition. New edition, revised and enlarged.

DOYLE A. & MCCRAE J. P. (2024). Developing a part-of-speech tagger for diplomatically edited Old Irish text. In R. SPRUGNOLI & M. PASSAROTTI, Éds., *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, p. 11–21, Torino, Italia: ELRA and ICCL.

DOYLE A., MCCRAE J. P. & DOWNEY C. (2019). A character-level LSTM network model for tokenizing the Old Irish text of the Würzburg glosses on the Pauline epistles. In *Proceedings of the Celtic Language Technology Workshop*, p. 70–79, Dublin, Ireland: European Association for Machine Translation.

FORKEL R., LIST J.-M., GREENHILL S. J., RZYMSKI C., BANK S., CYSOUW M., HAMMARSTRÖM H., HASPELMATH M., KAIPING G. A. & GRAY R. D. (2018). Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, **5**(1), 180205. DOI : 10.1038/sdata.2018.205.

FRANSEN T., ANDERSON C., BENIAMINE S. & PASSAROTTI M. (2024). The molor lemma bank: a new llod resource for old irish. In *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024*, p. 37–43, Torino, Italia: ELRA and ICCL.

GRIFFITH A., STIFTER D. & TONER G. (2018). Early Irish lexicography: a research survey. *Kratylos*, **63**, 1–28.

KAVANAGH S. & WODTKO D. S. (2001). *A lexicon of the Old Irish glosses in the Würzburg manuscript of the epistles of St. Paul*. Vienna: Verlag der Österreichischen Akademie der Wissenschaften.

KING D., LASH E. & GABAY L. (2006). In Dúil Bélrai: deilbhíocht an bhriathair. https://www2.smo.uhi.ac.uk/sengoidelc/duil-belrai/foirmeacha/. Implementation by Caoimhín P. Ó Donnaíle. Accessed: 2024-09-21.

LASH E., QIU F. & STIFTER D. (2020). *Morphosyntactic variation in medieval Celtic languages: corpus-based approaches*. Berlin, Boston: De Gruyter Mouton. DOI : 10.1515/9783110680744.

LURAGHI S., PALMERO APROSIO A., ZANCHI C. & GIULIANI M. (2024). Introducing PaVeDa – Pavia verbs database: valency patterns and pattern comparison in Ancient Indo-European languages. In R. SPRUGNOLI & M. PASSAROTTI, Éds., *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, p. 79–88, Torino, Italia: ELRA and ICCL.

MARSTRANDER C. *et al.*, Éds. (1913–1976). *Dictionary of the Irish Language: based mainly on Old and Middle Irish materials*. Dublin: Royal Irish Academy. 15 volumes.

MCCRAE J. P., BOSQUE-GIL J., GRACIA J., BUITELAAR P. & CIMIANO P. (2017). The OntoLex-Lemon model: development and applications. In *Proceedings of eLex 2017 conference*, p. 19–21.

NYHAN J. (2006). *The application of XML to the historical lexicography of Old, Middle and Early Modern Irish: a lexicon-based analysis*. Thèse de doctorat, University College Cork.

NYHAN J. (2008). Developing integrated editions of minority language dictionaries: the Irish example. *Literary and Linguistic Computing*, **23**(1), 3–12.

Ó CORRÁIN D., MORGAN H., FÄRBER B., HAZARD B., PURCELL E., Ó DÓNAILL C., LAVELLE H., NYHAN J. & MCCARTHY E. (1997). CELT: Corpus of Electronic Texts. https://celt.ucc.ie. Accessed: 2024-09-21.

Ó DONNAÍLE C. (2014). Tools facilitating better use of online dictionaries: technical aspects of Multidict, Wordlink and Clilstore. In J. JUDGE, T. LYNN, M. WARD & B. Ó RAGHALLAIGH, Éds., *Proceedings of the First Celtic Language Technology Workshop*, p. 18–27, Dublin, Ireland: Association for Computational Linguistics and Dublin City University. DOI : 10.3115/v1/W14-4603.

PASSAROTTI M., MAMBRINI F., FRANZINI G., CECCHINI F. M., LITTA E., MORETTI G., RUFFOLO P. & SPRUGNOLI R. (2020). Interlinking through lemmas. The lexical collection of the LiLa knowledge base of linguistic resources for Latin. *Studi e Saggi Linguistici*, **58**(1), 177–212.

PRUD'HOMMEAUX E. & SEABORNE A. (2008). *SPARQL Query Language for RDF*. W3C recommendation, W3C. https://www.w3.org/TR/rdf-sparql-query/. Accessed: 2024-09-21.

STIFTER D. (2009). Early Irish. In M. J. BALL & N. MÜLLER, Éds., *The Celtic languages*, p. 55–116. Routledge, 2nd édition.

STIFTER D., BAUER B., LASH E., QIU F., WHITE N., BARRETT S., GRIFFITH A., BULATOVAS R., FELICI F., GANLY E., NGUYEN T. H. & NOOIJ L. (2021). Corpus PalaeoHibernicum (CorPH) v1.0. https://chronhib.maynoothuniversity.ie. Accessed: 2024-09-21.

TONER G., NÍ MHAONAIGH M., ARBUTHNOT S., WODTKO D. & THEUERKAUF M.-L. (2019). Electronic Dictionary of the Irish Language. https://dil.ie. Accessed: 2024-09-21.

WILKINSON M. D., DUMONTIER M., AALBERSBERG I. J., APPLETON G., AXTON M., BAAK A. *et al.* (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, **3**(1), 1–9.