Ontologies for historical languages: using the LiLa and OntoLex-Lemon framework to build a Lemma Bank for Old Irish

Theodorus Fransen
CIRCSE Research Centre
Università Cattolica del Sacro Cuore
Largo A. Gemelli, 1
20123 Milan, Italy
theodorus.fransen@unicatt.it

Abstract

This paper presents a Linked Data approach to digitising and structuring Old Irish linguistic resources using the LiLa (Linking Latin) ontology, which is itself largely based on the OntoLex-Lemon framework (Cimiano et al., 2016). Old Irish, as a historical Celtic language with fragmented textual traditions, presents unique challenges for the creation and interoperability of digital resources. This work is part of the MOLOR project, whose aim is to create a knowledge base for Old Irish by interlinking texts, lexicons, and inflectional data. The first step in this ambitious endeavour is described here: the creation of an RDF linguistic Linked Data hub known as a Lemma Bank, similar to the one created as part of the LiLa project, addressing specific linguistic challenges and opportunities while adhering to the LiLa ontology.

1 Introduction

The digitisation of ancient and medieval languages presents significant challenges for computational linguistics and Digital Humanities scholars. Old Irish (600–900 CE) constitutes the earliest attested period of the Irish language—and of any Celtic language—for which the surviving documentary evidence is sufficiently comprehensive to allow a complete synchronic linguistic analysis; it is important for both the study of Indo-European linguistics and later stages of the Irish language (Stifter, 2009, 59). However, the combination of morphological complexity and orthographic variation (see Section 2.2), along with the use of different editorial standards, annotation schemas, and data formats in linguistic resources, creates substantial barriers to the successful use of standard natural language processing (NLP) methods (Doyle et al., 2019; Doyle and McCrae, 2024; Dereza et al., 2023). These factors also hinder resource compatibility and interoperability (Doyle and McCrae, 2025). The most important challenge in the current work is the variation seen in lemmatisation practice across lexical resources, particularly with regard to inflectional categorisation (see Section 4.1).

Recent advances in Linked Data technologies and semantic web frameworks offer promising solutions to address these challenges. The LiLa (Linking Latin) ontology, originally developed for Latin linguistic resources, provides a robust framework for representing ancient and historical language data in machine-readable formats that support interoperability and scholarly research (Passarotti et al., 2020).

This paper describes the implementation of the LiLa ontology—in turn adhering to OntoLex-Lemon—in the context of the Old Irish MOLOR project[1], presenting both methodological approaches and practical solutions to the unique challenges posed by this medieval Celtic language. The work contributes to the growing field of digital resources for ancient languages while demonstrating the adaptability of existing ontological frameworks to new linguistic contexts.

2 Background

2.1 Linked Data for ancient and historical languages

Chiarcos et al. (2018) address the underexplored application of Linked Open Data to ancient and historical languages and report on a case study applying LLOD principles to Assyriology by creating a Linked Data edition of the Electronic Text Corpus of Sumerian Royal Inscriptions. This linguistically annotated Sumerian corpus is connected to lexical resources, annotation terminology repositories, and museum collections housing the original cuneiform artifacts. The work serves as a foundation for expanding Linked Data approaches to other cuneiform corpora, including Ur III administrative

[1] https://tinyurl.com/molor-project

and legal texts, as part of the broader Machine Translation and Automated Analysis of Cuneiform Languages project (MTAAC, 2017–2020) and in close collaboration with the Cuneiform Digital Library Initiative or CDLI (CDLI contributors, 2025).[2]

Tittel and Chiarcos (2018) discuss the conversion of a medieval French medical treatise from a traditional scholarly edition into a semantically enriched digital format using RDFa (Adida et al., 2015) to link vocabulary entries to the Dictionnaire étymologique de l'ancien français (DEAF), offering a technological bridge between TEI/XML standards and Linked Open Data resources in digital philology.

The LiLa project[3] (Passarotti et al., 2020) represents a particularly relevant model, creating a comprehensive Linked Data ecosystem for Latin linguistic resources. LiLa's ontology provides sophisticated mechanisms for representing morphological, syntactic, and semantic information while maintaining interoperability across diverse resource types and scholarly traditions.

These projects demonstrate the feasibility and scholarly value of Linked Data approaches for historical language materials, enhancing discoverability, interoperability, and analytical capabilities.

2.2 Old Irish and linguistic challenges

Old Irish presents unique challenges for digital resource development that distinguish it from better-resourced ancient languages such as Latin or ancient Greek. Although Old Irish represents the first Celtic language with sufficient written evidence to enable comprehensive grammatical analysis, its associated contemporary text corpus—predominantly glosses on Latin manuscripts—is relatively small.

The language differs significantly from other Indo-European languages in several key ways. Its syntax follows a verb-first word order pattern (Stifter, 2009, 60), characteristic of Insular Celtic languages. Stifter (2009, 60) says the following about the linguistic complexity of Old Irish:

> Old Irish is almost prototypical for a language whose grammatical behaviour cannot be described adequately by synchronic rules. The bewildering complexities of some of its grammatical subsystems, especially that of verbal morphology, become transparent only when viewed from a diachronic position, and in order to understand allomorphic variation correctly it is essential to work with underlying forms and their often quite dissimilar surface representations

The same author continues with an illustrative example: "both *do·sluindi* /doˈslunʲdʲi/ '(s)he denies' and negated *ní·díltai* /ˈdʲiːlti/ '(s)he does not (*ní*) deny' regularly reflect the same diachronically underlying structure *dī-slondīθ" (Stifter, 2009, 60).

Phonologically, Old Irish has an extensive consonant inventory and displays what some scholars have termed a vertical vowel system (Anderson, 2016).[4] Although not unique to Old Irish, the language also exhibits initial mutations—changes to consonants and sometimes vowels at the start of the word based on grammatical context—whose orthographic encoding is neither systematic nor consistent in early texts. For example, a lenited *f* (which is silent) may be represented as *f*, *ḟ*, or may disappear altogether in the orthography.

These distinctive linguistic characteristics make the development of quality computational tools for Old Irish a pressing scholarly need.[5]

Old Irish orthography shows variation across manuscripts and time periods, reflecting both scribal practices and genuine linguistic change during the Old Irish period. Often, this orthographic variation is intertwined with morphological variation and change, creating additional challenges for automated processing and cross-referencing of textual materials and—most relevant for the purposes of the current paper—the selection and harmonisation of an exhaustive set of representative citation forms, i.e. lemmas, as illustrated in Section 4.1.

What may be viewed as orthographic or phonological variation may point to morphological variation, which may in turn be obscured by particular spellings. Taking the example of *cladaid* and *claidid* (see Appendix, Table 2), the difference here is the consonance (non-palatal vs palatal) of root-final *d* (i.e. non-palatal *clad-* vs palatal *claid-*), which

[2] The data structure of CDLI is illustrated at https://cdli.ox.ac.uk/wiki/doku.php?id=data_structure.

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

[4] A vertical vowel system in phonology refers to a vowel inventory organised primarily along the height dimension (high, mid, low) with minimal or no distinctions based on frontness/backness or rounding.

[5] For a more extensive discussion on the morphological complexity of the Old Irish verb in the context of computational modelling see Fransen (2020).

signifies a difference in inflection class. An orthographic representation such as *cladid*, however, could represent either morphological variant.

The fragmentary nature of the Old Irish corpus also creates challenges for comprehensive coverage, as many forms and morphonological contexts are only attested rarely or in ambiguous contexts. This requires careful balance between exhaustive representation and practical utility in ontological design decisions. The next subsection discusses the lexicographical landscape and lists the resources instrumental for compiling an Old Irish Lemma Bank.

2.3 The Old Irish resource landscape

Griffith et al. (2018) give an overview of lexicographical resources available for early medieval Irish. Dereza (2018) is a first valuable and instructive attempt at building a lemmatiser for Old Irish using rule-based and machine learning techniques (based on DIL, see below).

Stifter et al. (2022) call for greater interoperability between linguistic resources for early medieval Irish. Indeed, there has recently been a push towards resource interoperability and standardisation. Doyle and McCrae (2025) report on a new lexical resource and the publication of two treebanks following the Universal Dependencies (de Marneffe et al., 2021) standard of annotation, noting that these resources "have been created with the express purpose of ensuring lexical compatibility between them" (p. 393). The equally novel resource Goidelex (Anderson et al., 2024) incorporates normalised orthographical forms and is compatible with other frameworks (see below).

Despite these promising developments, resources currently do not speak the same language, i.e. there is a lack of a unified ontology and controlled vocabularies following Semantic Web standards. The current work is a first step in overcoming this limitation, by creating a collection of canonical forms, i.e. lemmas, used to interlink resources using Linguistic Linked Data methods following the LiLa framework (Passarotti et al., 2020). The current work uses the following three resources for the collection of lemmas.

Dictionary of the Irish Language (DIL) (eDIL, 2019)—The standard dictionary for medieval Irish covering the period 700–1700CE, which transitioned from print to digital format in 2007. While DIL offers extensive lexical coverage, it suffers from non-exhaustive and inconsistent annotation of examples and limited data extraction functionalities for large-scale research. Furthermore, headwords are not always representative of Old Irish.

Corpus PalaeoHibernicum (CorPH) (Stifter et al., 2021)—CorPH constitutes the most morphosyntactically detailed and comprehensive lexical resource for Old Irish. It contains over 10,500 word entries from 77 analysed texts, available as downloadable CSV files. While not immediately relevant for the task of building a Lemma Bank, its complex word structure breakdown makes it difficult to link back to source texts.

Würzburg lexicon (Kavanagh and Wodtko, 2001)—A print dictionary accompanied by PDF files for the highly important 8th-century Würzburg glosses (not covered in CorPH).

Goidelex (Anderson et al., 2024)—This novel resource currently contains 671 entries from the Würzburg glosses, extracted from Kavanagh and Wodtko (2001). It provides detailed inflectional and phonological data, uses normalised spelling, links to other resources, and follows modern data standards, including Paralex, a novel standard for inflectional lexicons (Beniamine et al., 2023). Paralex includes tools for converting data into the RDF OntoLex-Lemon format. It is also compatible with the Cross Linguistic Data Format (Forkel et al., 2018).

3 Modelling: the LiLa lemma ontology

The LiLa ontology provides a comprehensive framework for representing linguistic data through Linked Data principles. The LiLa knowledge base centres on a comprehensive collection of Latin lemmas that serve as connection points between different language resources. Since the system is vocabulary-focused, these lemmas link together dictionary entries, corpus texts and NLP output that reference the 'common denominator', enabling seamless integration across resources (Passarotti et al., 2020, 186–187). The ontology incorporates multiple levels of linguistic analysis, from graphemic representation through morphological, lexical, and syntactic annotation to semantic and pragmatic information. The ontology employs standardised vocabularies and URI schemes that enable cross-referencing between different resources and projects, supporting both human-readable scholarly annotation and machine-processable data that can be queried and analysed computationally. The LiLa Lemma ontology is described and exemplified in

Listing 1: LiLa Lemma Class Definition

```
lila:Lemma a rdfs:Class,
             owl:Class ;
    rdfs:comment "A Lemma must have a POS, but it cannot have more than 1",
                 "In LiLa, a Lemma is a form in the word inflection that is used (or may
     potentially be used) to lemmatize tokens in a corpus." ;
    rdfs:label "Lemma" ;
    rdfs:subClassOf ontolex:Form ;
    rdfs:subClassOf [ a owl:Restriction ;
                      owl:onClass lila:POS ;
                      owl:onProperty lila:hasPOS ;
                      owl:qualifiedCardinality "1"^^xsd:nonNegativeInteger
                    ] .
```

Listing 2: The entry *sequor* 'to follow' as modelled according to the LiLa Lemma Class

```
<data/id/lemma/124461>
        a                       lila:Lemma ;
        rdfs:label              "sequor" ;
        lila:hasBase            <data/id/base/417> ;
        lila:hasInflectionType  lila:v3d ;
        lila:hasPOS             lila:verb ;
        lila:lemmaVariant       <data/id/lemma/124462> ;
        dcterms:isPartOf        <data/id/lemma/LemmaBank> ;
        ontolex:writtenRep      "sequor"@la , "secor"@la .
```

Listings 1 and 2, respectively.

Although a discussion on the application of LiLa and OntoLex-Lemon classes and properties to Old Irish lexemes has already been provided in Fransen et al. (2024), it might be prudent to briefly explain the `lila:lemmaVariant` property here again. This property was created in LiLa to cater for the use of alternative canonical forms used for the same lexeme as represented in lexical entries while at the same time maintaining resource interoperability. Consider Figure 1. Here we see four inflectional variants—first vs second conjugation, active vs deponent—representing four citation forms for the same Latin lexeme 'to limp'. Using the commutative property `lemmaVariant`, "LiLa harmonises different lemmatisation strategies and annotation styles, thus granting interoperability" (Passarotti et al., 2020, 193). This is exactly because each resource or token linked to one of those lemmas is linked to any other token or resource lemmatised using one of the variant lemmas. Note that this elegantly circumvents the restriction that an `ontolex:LexicalEntry` can have at most one canonical form (Cimiano et al., 2016, §3.1).

For purely orthographic variation (or certain phonological variants in the case of Old Irish (Fransen et al., 2024)), the different spellings are modelled according to the more general property `representation` (`ontolex:WrittenRep`) and, crucially, *as part of the same lemma and hence URI*—compare the written variants *sequor* and *secor* in Listing 2 and *claudo* and *cludo* in Figure 1, respectively.

For the Old Irish implementation, LiLa's core concepts have been adhered to—not without challenges—as detailed in Section 4.1.

4 Implementation

4.1 Harmonisation challenges

Adapting the LiLa ontology for Old Irish presents several significant challenges that require careful methodological consideration. As discussed in Section 2.2, Old Irish is characterised by a high degree of synchronically unpredictable (or at least opaque) allomorphy, and in this respect arguably exceeds the morphological complexity of Latin, particularly in verbal inflection.[6]

Admittedly, a high degree of allomorphy, combined with spelling variation, is not necessarily problematic for the task of collecting and aligning lemmas from already existing lexical resources, as

[6] Although few Celtic and classical scholars would disagree, the author is not aware of any empirical study that compares Old Irish and other historical Indo-European languages such as Latin using features of morphological complexity; however, the reader may want to consult Fransen (2019, 30–34) for some quantitative observations on the Old Irish verbal system.

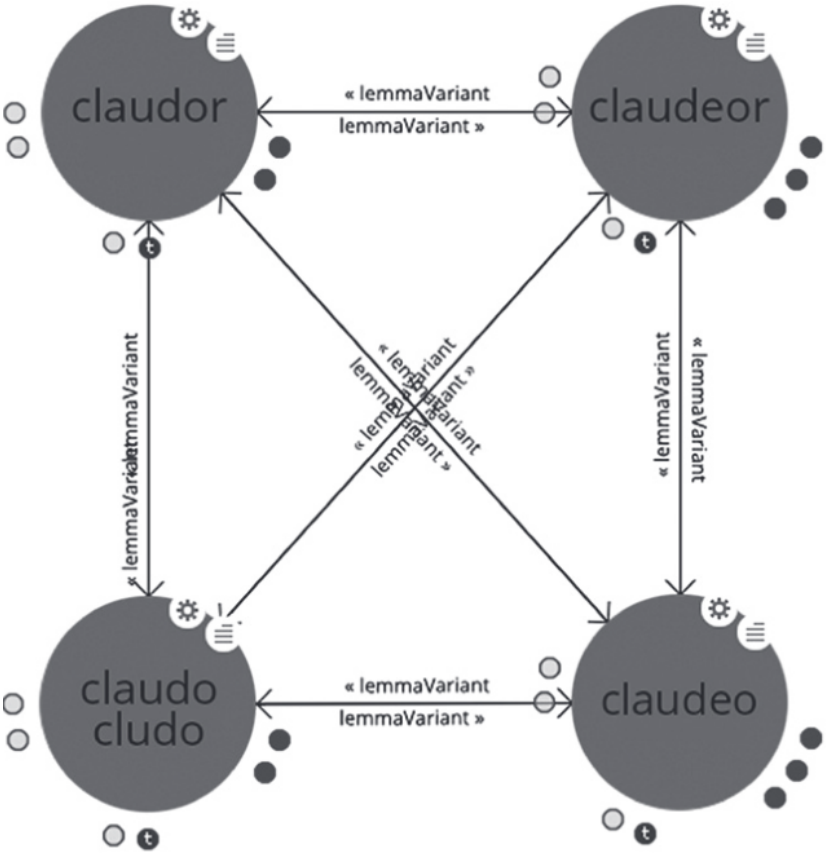

Figure 1: Four Latin lemmas with different inflection patterns representing different citation forms, connected through the commutative property `lila:lemmaVariant`; taken from Passarotti et al. (2020, 193).

described in this paper (as opposed to automatic morphological analysis and lemmatisation of raw text). The challenge at hand, more so than morphological complexity itself, has proven to be the lack of uniformity between resources in the categorisation of morphological (or more specifically, inflectional) variation. Linguistic complexity and variation and lack of descriptive uniformity are obviously related, with other factors at play, such as uncertainty due to gaps in attestation.

Inflectional variation is the pillar of the lemma variant property. Notwithstanding this property's usefulness, mapping the variation seen in Old Irish data onto clear-cut inflectional variants proved rather challenging, especially with the nominal system. Tables 1 and 2 in the Appendix provide an overview of the inflectional variation and microclasses seen with Old Irish lemmas. Notably in the case of the nominal system, one can observe 1) unbalanced categorisation of lemmas (one-to-many relationships) across resources; 2) different, yet sometimes overlapping inflection classes (`fl_cat`); and 3) different resolution (i.e. macro- vs. microclasses). The noun *fius* 'knowledge', for example, is described as a *u*-stem or *o*-stem, and can be masculine or neuter. Creating four lemmas for all four permutations seems excessive and is not reflective of Old Irish—the linguistic reality is a mixed (and not fully attested) inflectional paradigm due to confusion between stems and a general shift of *u*-stems towards *o*-stem inflection (Thurneysen, 1946, §309).

Furthermore, as can be partially gleaned from the footnotes accompanying the tables in the Appendix, the annotation of inflectional classes is at times arbitrary, inconsistent, or incorrectly suggestive of differences; a good example of the latter is the inclusion of both the lemmas *bádaid* and *báidid* 'to drown' in CorPH—see Table 2, footnote *c*. All this is compounded by the occasional slip in CorPH.[7]

Orthographic variation presents another significant challenge, as Old Irish manuscripts show considerable spelling variation between different scribes and across chronological periods. The on-

[7] As part of the data extraction process and resource alignment, the author has already identified and corrected hundreds of mistakes in CorPH, ranging from typos to the assignment of the wrong language to an entry.

tology must support multiple orthographic representations while maintaining scholarly precision in distinguishing between genuine linguistic (mostly inflectional) variation and scribal variation and inconsistency. The written representation data property of `ontoLex:Form`, like in LiLa, was considered sufficient to encode orthographic variation (as opposed to encoding morphological variation, for which the `lila:lemmaVariant` property, as detailed in Section 3, was used).

That being said, the primary purpose of a Lemma Bank is to accommodate and unify variation found in lemmatisation practice in order to make lexical resources interoperable; it is not the place for prioritising certain spellings or providing a highly principled and systematic morphological categorisation of forms. Of course, a Lemma Bank can be a first point of call and as such might benefit from linguistic means that facilitate search queries, linguistic description, or research purposes (e.g. using typographical means consistently to make compounding more explicit, see Section 4.2), as long as the matching of lemmas with lexical entries (or perhaps even tokens in a text) remains computationally trivial.

Since it contains lemmas from existing resources, a Lemma Bank naturally inherits some of the lemmatisation inconsistencies found in those resources. However, by means of 1) exhaustive coverage of (potential) lemmas and 2) the principles of Linked Data, and especially the SPARQL query language (Prud'hommeaux and Seaborne, 2008), interoperability and effective retrieval of linguistic information in linked specialist resources, possibly built with standardisation or normalisation in mind, is warranted.

Goidelex (Anderson et al., 2024), which employs a normalised orthography for Old Irish (Fransen et al., 2023), may serve as an example. Since it is built according to the Paralex standard for inflected lexicons, which, as mentioned earlier, includes an ontology for conversion into OntoLex-Lemon lexicons, there exists the theoretical possibility of navigating from a lemma in the Lemma Bank to the associated lexical entry in Goidelex and retrieving inflectional paradigms in normalised orthography.

4.2 SQL tables

RDF conversion was preceded by semi-automatic creation and integration of SQL tables based on the extraction of data from the resources mentioned in Section 2.3, currently limited to nouns (including verbal nouns and proper nouns), numerals, and verbs.

For nouns, the author integrated CorPH's lemma table, a selection of compositional forms entries from CorPH's morphology table, and Goidelex and Kavanagh and Wodtko (2001)[8] (for the Würzburg lemmas, which are not in CorPH). For verbs, CorPH was again used, manually aligned with the verbal subset in CSV files extracted from the Würzburg lexicon (Kavanagh and Wodtko, 2001)[9] and verb entries from DIL. The mid-high dot has been invariably employed with compound verbs (rather than the hyphen, as used in, e.g, DIL), separating the pretonic preverb from the stressed part of the verb, e.g. *do·beir* 'to give, bring'. Compound nouns were hyphenated broadly following Kavanagh and Wodtko (2001), even where they were not hyphenated in other source data, e.g. *dag-athair* 'good father', primarily with a view to creating typographical consistency among lemmas.[10] Table 2 in the Appendix closely mirrors (a snippet of) the initial spreadsheet (apart from the URIs, of course) used to manually align verb lemmas—subsequently converted into a TSV file and imported as an SQL table.

4.3 RDF conversion

The Old Irish lemma data in the relational databases—*lemma*, *lemma_wr* and *variant_group*—was subsequently converted into RDF using the D2RQ mapping language (Cyganiak et al., 2012), emulating the URI schemes for the LiLa Lemma Bank. However, at least in the first instance, fewer properties have been used, the absence

[8] More precisely, the Goidelex lexemes table which represents (normalised) entries with more than one attestation, plus hapax lemmas manually added from the Würzburg lexicon.

[9] The lexicon was automatically parsed and converted into CSV files by Dr Aaron Griffith on the basis of accompanying PDF files, with assistance from the Utrecht Digital Humanities Lab. The parsing script is found at `https://github.com/CentreForDigitalHumanities/wurzburg-glosses-extraction` while the CSV files were generously shared privately with the author. Admittedly, the parsed files only cover a selection of POS categories and some entries are missing (some verbs had to be manually added). Moreover, the extraction is noisy in places.

[10] These typographical separation devices reflect morphological boundary markers which are linguistically insightful, even though their inclusion might arguably go beyond the remit of a Lemma Bank. Having said this, ignoring these markers in queries is trivial, while inserting them post hoc is not. Furthermore, they can easily be deleted in a string manipulation step to facilitate matching with linguistic resources that do not employ these markers, such as diplomatically edited text resources.

Listing 3: The form *breth* 'bearing' as modelled according to the MOLOR Lemma Class

```
<http://molor.eu/data/id/lemma/1490>
      a       molor:Lemma ;
      rdfs:label "breth" ;
      molor:hasPOS molor:noun ;
      molor:lemmaVariant <http://molor.eu/data/id/lemma/4924> ;
      ontolex:writtenRep "breth" .
```

Listing 4: The form *brith* 'bearing' as modelled according to the MOLOR Lemma Class

```
<http://molor.eu/data/id/lemma/4924>
      a       molor:Lemma ;
      rdfs:label "brith" ;
      molor:hasPOS molor:noun ;
      molor:lemmaVariant <http://molor.eu/data/id/lemma/1490> ;
      ontolex:writtenRep "brith" .
```

of `lila:hasInflectionType` probably being the most significant difference (see Section 5). The Lemma Bank currently totals 6,000+ lemmas, a fifth of which are verbs.

The URI schemes otherwise follow LiLa conventions, ensuring future compatibility with ancient and historical language Linked Data resources. Listings 3–6 exemplify the RDF version of the entries for the nouns *breth* and *brith* 'bearing' as well as for the verbs *molaithir* (deponent) and *molaid* (active) 'to praise', illustrating the author's choices in employing the written representation datatype vs the lemma variant property with these forms (the reader may want to refer to Tables 1 and 2 in the Appendix, respectively, for more details).

5 Discussion

5.1 Recapitulation: scope and function of a Lemma Bank

Inconsistent or divergent annotation of inflection types has presented the most complex aspect of the collecting and modelling lemmas from legacy resources, as discussed in Section 4.1. It was decided to not try and facilitate divergent inflectional annotation practices as part of the MOLOR RDF Lemma Bank, as this would have entailed having to focus on the linguistic exercise of (further) correcting and harmonising annotation in existing resources, which would most likely have meant choosing one categorisation system over another. Echoing what was discussed in Section 4.1, the goal of a Lemma Bank is to capture variant lemmatisation practices rather than aiming for standardisation and normalisation. Moreover, taking a principled and fine-grained approach to morphological variation would redundantly emulate work as part of Goidelex (Anderson et al., 2024), which is focused on providing high-resolution inflectional information employing a normalised orthography. Furthermore, considering the fact that a `molor:Lemma` (and `lila:Lemma`) is a subclass of `ontolex:Form`, the absence of morphological information is actually in line with the OntoLex-Lemon core model, where it is the lexical entry that is assigned morphological properties and not the form.

More generally, leaving specialised information to individual resources conforms to the philosophy of the Linked Data paradigm and its premise of knowledge being distributed (even if potentially divergent or conflicting in nature).

5.2 Applications to computational tasks

Despite preserving variation rather than enforcing standardisation, the harmonised lemma representations in the RDF Lemma Bank could significantly assist computational lemmatisation efforts for Old Irish. Work such as Dereza (2018) demonstrates the challenges of automatic lemmatisation for historical languages, where morphological complexity and orthographic variation create substantial obstacles. By providing a comprehensive, structured repository of lemma-form relationships across multiple lexical resources, the Lemma Bank offers a rich training resource that could improve lemmatisation accuracy. The ontological structure allows for sophisticated querying of lemma variants and their attestations, potentially enabling more robust handling of the orthographic and morphological variation that characterises Old Irish texts.

Listing 5: The form *molaithir* 'to praise' as modelled according to the MOLOR Lemma Class

```
<http://molor.eu/data/id/lemma/5744>
      a       molor:Lemma ;
      rdfs:label "molaithir" ;
      molor:hasPOS molor:verb ;
      molor:lemmaVariant <http://molor.eu/data/id/lemma/5745> ;
      ontolex:writtenRep "molaidir" , "molaithir" .
```

Listing 6: The form *molaid* 'to praise' as modelled according to the MOLOR Lemma Class

```
<http://molor.eu/data/id/lemma/5745>
      a       molor:Lemma ;
      rdfs:label "molaid" ;
      molor:hasPOS molor:verb ;
      molor:lemmaVariant <http://molor.eu/data/id/lemma/5744> ;
      ontolex:writtenRep "molaid" .
```

5.3 Advantages of RDF over traditional data formats

The choice of RDF over traditional relational databases or flat file formats (TSV, CSV) reflects the distributed and interconnected nature of lexical knowledge. While SQL databases excel at structured queries within closed systems, RDF graphs enable seamless integration across heterogeneous resources and institutions. This is particularly valuable for historical linguistics, where lexical data often originates from multiple scholarly traditions and projects. The graph-based model naturally represents the complex relationships between lemmas, forms, and attestations, while SPARQL queries can traverse these relationships in ways that would require complex joins in relational systems. Moreover, the use of standardised vocabularies like OntoLex-Lemon ensures interoperability with other linked lexical resources, facilitating broader comparative and cross-linguistic research that would be challenging to achieve with isolated database systems.

6 Conclusion and future work

The current work has focused on building an RDF Lemma Bank for Old Irish to interconnect linguistic resources according to semantic web principles and the LiLa ontology in particular. The application of the LiLa ontology to Old Irish demonstrates both the potential and challenges of existing Linked Data frameworks for under-resourced ancient and historical languages. A clear-cut mapping to LiLa lemma properties is not always trivial due to morphological and orthographic variation, inconsistency, or different resolution in inflectional annotation (to which we can add uncertainty originating in gaps in attestation). The decision was made not to enforce a single, harmonised morphological annotation system within the Lemma Bank, but instead to leverage the `lila:lemmaVariant` property (currently without using `lila:hasInflectionType`) to interlink alternative lemmas and `ontolex:writtenRep` for orthographic variation, thus respecting the distributed nature of Linked Data.

The Lemma Bank is expected to grow in size (more lemmas and more POS categories), with the linking of resources to the Lemma Bank constituting the beginnings of a knowledge base for Old Irish, with SPARQL endpoints that support complex morphological and syntactic searches throughout the Old Irish corpus.[11] Such a knowledge base will hopefully lead to enhanced search and analysis capabilities, while also highlighting areas where traditional philological approaches remain necessary supplements to computational methods.

User feedback from the scholarly community will be the best measure of project success; the author hopes to report on use cases during a future edition of the OntoLex workshop series.

Acknowledgments

MOLOR—Morphologically Linked Old Irish Resource—has received funding as part of the European Union's Horizon Europe scientific research initiative under the Marie Skłodowska-Curie Ac-

[11] The author is particularly interested in making two lexical resources interoperable with the Lemma Bank: Goidelex (Anderson et al., 2024) and the Irish section of PaVeDa (Roma and Zanchi, 2025), the latter of which currently consists of some hundred Old Irish verb entries marked for valency patterns.

tions (MSCA), grant agreement No 101106220. The author thanks the (former) LiLa project team for their generous sharing of expertise and resources. The initial indexing of verb lemmas was performed by Federico Simone Samperi (Università di Pavia). The author also acknowledges the contributions of Celtic Studies scholars, especially Prof. Elisa Roma, who has been kindly providing linguistic assistance over the course of the MOLOR project.

References

Ben Adida, Mark Birbeck, Shane McCarron, and Steven Pemberton. 2015. RDFa core 1.1 - third edition. W3c recommendation, World Wide Web Consortium.

Cormac Anderson. 2016. *Consonant colour and vocalism in the history of Irish*. Ph.D. thesis, Adam Mickiewicz University, Poznań.

Cormac Anderson, Sacha Beniamine, and Theodorus Fransen. 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*, pages 1–10, Torino, Italia. ELRA and ICCL.

Sacha Beniamine, Cormac Anderson, Mae Carroll, Matías Guzmán Naranjo, Borja Herce, Matteo Pellegrini, Erich Round, Helen Sims-Williams, and Tiago Tresoldi. 2023. Paralex: a DeAR standard for rich lexicons of inflected forms. In *International Symposium of Morphology*. https://www.paralex-standard.org.

CDLI contributors. 2025. About CDLI. https://cdli.earth/about. [Online; accessed 2025-07-13].

Christian Chiarcos, Émilie Pagé-Perron, Ilya Khait, Niko Schenk, and Lucas Reckling. 2018. Towards a linked open data edition of sumerian corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), May 7–12, 2018, Miyazaki, Japan*, pages 2437 – 2444.

Philipp Cimiano, John P. McCrae, and Paul Buitelaar. 2016. Lexicon Model for Ontologies: Community report. W3C community group final report, World Wide Web Consortium. https://www.w3.org/2016/05/ontolex/.

Richard Cyganiak, Chris Bizer, Jörg Garbers, Oliver Maresch, and Christian Becker. 2012. The D2RQ mapping language. v0.8 – 2012-03-12.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Oksana Dereza. 2018. Lemmatization for ancient languages: Rules or neural networks? In *Artificial Intelligence and Natural Language*, pages 35–47, Cham. Springer International Publishing.

Oksana Dereza, Theodorus Fransen, and John P. Mccrae. 2023. Do not trust the experts - how the lack of standard complicates NLP for historical Irish. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 82–87, Dubrovnik, Croatia. Association for Computational Linguistics.

Adrian Doyle and John McCrae. 2025. Development of Old Irish lexical resources, and two Universal Dependencies treebanks for diplomatically edited Old Irish text. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 393–402, Albuquerque, USA. Association for Computational Linguistics.

Adrian Doyle and John P. McCrae. 2024. Developing a part-of-speech tagger for diplomatically edited Old Irish text. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 11–21, Torino, Italia. ELRA and ICCL.

Adrian Doyle, John P. McCrae, and Clodagh Downey. 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*, pages 70–79, Dublin, Ireland. European Association for Machine Translation.

eDIL. 2019. An electronic dictionary of the Irish language. Based on the Contributions to a Dictionary of the Irish Language (Dublin: Royal Irish Academy, 1913–1976).

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and reuse in comparative linguistics. *Scientific Data*, 5(180205).

Theodorus Fransen. 2019. *Past, present and future: Computational approaches to mapping historical Irish cognate verb forms*. Ph.D. thesis, Trinity College Dublin, Dublin.

Theodorus Fransen. 2020. 3 Automatic morphological analysis and interlinking of historical Irish cognate verb forms. In Elliott Lash, Fangzhe Qiu, and David Stifter, editors, *Morphosyntactic Variation in Medieval Celtic Languages. Corpus-Based Approaches*, chapter 3, pages 49–84. De Gruyter Mouton, Berlin, Boston.

Theodorus Fransen, Cormac Anderson, and Sacha Beniamine. 2023. Towards a normalised orthography for Old Irish. Paper at *36th Irish Congress of Medievalists*, Dublin, 22–23 June 2023.

Theodorus Fransen, Cormac Anderson, Sacha Beniamine, and Marco Passarotti. 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*, pages 37–43, Torino, Italia. ELRA and ICCL.

Aaron Griffith, David Stifter, and Gregory Toner. 2018. Early Irish lexicography – a research survey. *Kratylos*, 63(1):1–28.

Séamus Kavanagh and Dagmar S. Wodtko. 2001. *A lexicon of the Old Irish glosses in the Würzburg manuscript of the epistles of St. Paul*. Verlag der Österreichischen Akademie der Wissenschaften, Vienna.

Kim McCone. 1987. *The Early Irish verb*. An Sagart, Maynooth.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 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.

Eric Prud'hommeaux and Andy Seaborne. 2008. SPARQL query language for RDF.

Helmut Rix and Martin Kümmel. 2001. *LIV, Lexikon der indogermanischen Verben: die Wurzeln und ihre Primärstammbildungen*, 2., erw. und verb. aufl. edition. Reichert, Wiesbaden.

Elisa Roma and Chiara Zanchi. 2025. Old Irish in the PaVeDa: Issues, perspectives, and two case studies. In Dylan R. Cooper, Rachel Martin, Graham O'Toole, and Samuel Ezra Puopolo, editors, *Proceedings of the Harvard Celtic Colloquium 42: 2023*, volume 42, pages 212–237. Harvard University Press, Boston.

David Stifter. 2009. Early Irish. In Martin Ball and Nicole Müller, editors, *The Celtic Languages*. Routledge.

David Stifter, Bernhard Bauer, Elliott Lash, Fangzhe Qiu, Nora White, Siobhán Barrett, Aaron Griffith, Romanas Bulatovas, Francesco Felici, Ellen Ganly, Truc Ha Nguyen, and Lars Nooij. 2021. Corpus PalaeoHibernicum (CorPH) v1.0. `https://chronhib.maynoothuniversity.ie`.

David Stifter, Nina Cnockaert-Guillou, Beatrix Färber, Deborah Hayden, Máire Ní Mhaonaigh, Joanna Tucker, and Christopher Guy Yocum. 2022. Developing a digital framework for the medieval Gaelic world. Project report, Queen's University Belfast.

Rudolf Thurneysen. 1946. *A Grammar of Old Irish*. Dublin Institute of Advanced Studies, Dublin. Translated by D. A. Binchy and Osborn Bergin.

Sabine Tittel and Christian Chiarcos. 2018. Historical lexicography of Old French and linked open data: transforming the resources of the dictionnaire étymologique de l'ancien français with OntoLex-Lemon. In *Proceedings of the 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science, co-located with LREC2018, 12 May 2018, Miyazaki, Japan*.

URI	CorPH					Goidelex					DIL[a]					Meaning
	ID	Entry	POS	fl_cat	Gender	ID	Entry	POS	fl_cat	Gender	ID	Entry	POS	fl_cat	Gender	
http://molor.eu/data/id/lemma/4670	4627	*fius*	noun	u	n	fius-282	*fius*	noun	u	n	dil.ie/22221	1 *fis, fius*	noun	u, o	n, m	know-ledge[b]
	—	—	—	—	—	fius-282-1	*fius*	noun	u	m	—	—		—	—	
	—	—	—	—	—	fius-282-2	*fius*	noun	o	n	—	—	—	—	—	
—	—	—	—	—	—	—	—	—	—	—	dil.ie/21774	*fes(s)*[c]	x[d]	—	—	
http://molor.eu/data/id/lemma/1490	430	*breth*	verbal_noun	ā	f	—	—	—	—	—	dil.ie/6785	*breth*	verbal_noun	ā, i	f	bearing[e]
http://molor.eu/data/id/lemma/4924	9582	*brith*	verbal_noun	i[f]	f	brith-97	*brith*	verbal_noun	i	f	dil.ie/6860	*brith*	x	—	—	
—	—	—	—	—	—	breith-97-1	*breith*	verbal_noun	ā	f	dil.ie/6698	*breith*	x	—	—	
http://molor.eu/data/id/lemma/4410	5230	*adaig*	noun	ī	f	adaig-524	*adaig*	noun	n1[g]	f	dil.ie/256	1 *adaig*	noun	iā	f	night
—	—	—	—	—	—	—	—	—	—	—	dil.ie/856	*aidche*	x	—	—	
—	—	—	—	—	—	—	—	—	—	—	dil.ie/33617	*oidche*	x	—	—	
http://molor.eu/data/id/lemma/4935	3261	*pendaind*	noun	ī	f	—	—	—	—	—	dil.ie/34272	*pennaind*	noun	—	f	penance
http://molor.eu/data/id/lemma/4807	3262	*pendait*	noun	ī	f	pennait-362	*pennait*	noun	ī2	f	dil.ie/34273	*pennait*	noun	ā	f	
http://molor.eu/data/id/lemma/4638	5506	*eipistil*	noun	ā/i/ī	f	eipistil-551	*eipistil*	noun	ī2	f	dil.ie/20187	*epistil*	noun	—	f	epistle, letter
http://molor.eu/data/id/lemma/4873	8894	*talam*	noun	n	m	talam-889	*talam*	noun	n1	m	dil.ie/39932	*talam*	noun	n	m	earth, land
http://molor.eu/data/id/lemma/4708	7502	*gein*	noun	n	n	gein-750	*gein*	verbal_noun	n2	n	dil.ie/25530	1 *gein*	verbal_noun, noun	—	n, f	the act of pro-cre-ation, birth
http://molor.eu/data/id/lemma/4808	6739	*persan*	noun	ā/n	f	persan-673	*persan*	noun	ā	f	dil.ie/34285	*persa*	noun	n	f	person

Table 1: Comparison of Old Irish nominal lexical entries across CorPH, Goidelex, and DIL resources

[a] Nouns have not yet been systematically aligned with DIL

[b] Also used as the verbal noun of the verb *ro·finnadar* 'to find out'.

[c] Neuter plural form.

[d] The *x* denotes a cross-reference to the main DIL entry.

[e] Also as common noun with meaning 'judgment', as part of the same entry in DIL.

[f] This verbal noun has a gen.sg in *-e* and as such does not fully adhere to *i*-stem inflection, even though this stem designation may be etymologically correct; see Thurneysen (1946, §§256, 294) on the blurred lines between *ā*- and *i*-stem inflection with certain verbal nouns.

[g] For a description of inflectional microclasses see https://github.com/cormacanderson/Goidelex/blob/main/inherent_properties.csv.

URI	CorPH				Kavanagh		DIL				Meaning
	ID	Entry	POS	fl_cat	Entry	POS	ID	Entry	POS	fl_cat	
http://molor.eu/data/id/lemma/6097	3165	*ad·roilli*	verb	H2[a]	*ad·roilli*	verb	dil.ie/558	*ad·roilli*	verb	—	to deserve
http://molor.eu/data/id/lemma/6098	6153	*as·roilli*	verb	H2	—	—	dil.ie/4482	*as·roilli*	x[b]	—	
http://molor.eu/data/id/lemma/6087	318	*báidid*	verb	W2b[c]	—	—	dil.ie/5172	*báidid*	verb	—	to submerge
	3517	*bádaid*	verb	W2a	—	—	—	—	—	—	
http://molor.eu/data/id/lemma/6170	3711	*cladaid*	verb	S1a	—	—	dil.ie/9297	*cladaid*	x	—	to dig
http://molor.eu/data/id/lemma/6171	8275	*claidid*	verb	S2	*claidid*	verb	dil.ie/9329	*claidid*	verb	—	
http://molor.eu/data/id/lemma/5744	5395	*molathair*[d]	verb	W1	*molaidir*	verb	dil.ie/50393	*molaithir*	x	—	
http://molor.eu/data/id/lemma/5745	—	—	—	—	*molaid*	verb	dil.ie/32491	*molaid*	verb	ā	to praise
http://molor.eu/data/id/lemma/5906	5893	*taraisnigidir*	verb	W2a	—	—	dil.ie/39730	*tairisnigidir*	verb	g depon	to trust in
	6012	*toraisnigidir*	verb	W2a	—	—	—	—	—	—	

Table 2: Comparison of Old Irish verbal entries across CorPH, Kavanagh, and DIL resources

[a] The classification system used here is from McCone (1987); another widely used classification is Thurneysen (1946).

[b] The *x* denotes a cross-reference to the main DIL entry.

[c] This verb derives from the PIE causative formation $*g^{u}oh_2d^{h}$-*éi̯e*- (Rix and Kümmel, 2001, 206), which explains the class type W2b, reserved for causatives with (mostly) an *o* or *u* as root vowel (McCone, 1987, 28). Synchronically, however, the distinction between W2a and W2b is of little relevance with this verb—albeit confusingly suggestive of a difference—as it has no bearing on either the conjugation pattern or the neutral vs palatal consonance; as such, these lemmas have been merged, each having been assigned a `ontolex:writtenRep` as part of the same URI.

[d] This should be *molaithir* (Prof. David Stifter (Maynooth University), pers. comm.) and has been corrected accordingly; see also Thurneysen (1946, §575).