Goidelex: A Lexical Resource for Old Irish

Cormac Anderson[†], **Sacha Beniamine**[†], **Theodorus Fransen**[‡]
[†]University of Surrey, Guildford, United Kingdom
[‡]Università Cattolica del Sacro Cuore, Milan, Italy
{cormac.anderson, s.beniamine}@surrey.ac.uk, theodorus.fransen@unicatt.it

Abstract

We introduce Goidelex, a new lexical database resource for Old Irish. Goidelex is an openly accessible relational database in CSV format, linked by formal relationships. The launch version documents 695 headwords with extensive linguistic annotations, including orthographic forms using a normalised orthography, automatically generated phonemic transcriptions, and information about morphosyntactic features, such as gender, inflectional class, etc. Metadata in JSON format, following the Frictionless standard, provides detailed descriptions of the tables and dataset. The database is designed to be fully compatible with the Paralex and CLDF standards and is interoperable with existing lexical resources for Old Irish such as CorPH and DIL. It is suited to both qualitative and quantitative investigation into Old Irish morphology and lexicon, as well as to comparative research. This paper outlines the creation process, rationale, and resulting structure of the database.

Keywords: Old Irish, morphology, lexicon, inflection

1. Introduction

We present Goidelex,[1] a new lexical database of Old Irish[2] which draws on and adds to existing digital resources for the language. The launch version of the database documents 695 headwords, in both orthographic forms and phonemic transcription and with extensive linguistic annotation. It is structured and formatted as a set of CSV files and is designed to be forward compatible with the Paralex and CLDF standards (Beniamine et al., 2023; Forkel et al., 2018). While the launch dataset contains only nominal lexemes, the database has been designed with a view to adding also other parts of speech in the near future. As a standalone resource, Goidelex is suited for both qualitative and quantitative investigation of the Old Irish lexicon. Moreover, it links between several existing resources: the electronic Dictionary of the Irish Language (DIL: Toner et al. (2013-present)), the Corpus PalaeoHibernicum (CorPH: Stifter et al. (2021)) and the Würzburg glosses (Kavanagh and Wodtko, 2001), facilitating research on Old Irish phonology and morphology.

In recent times, computational methods have increasingly come to be used to investigate various aspects of linguistic typology and evolution. However, these methods require well-structured machine-readable data, which is most often available only for well-resourced, literary languages with lots of speakers (Dahl, 2015; Malouf et al., 2020; Bird, 2022). This unbalanced sampling makes it especially important to develop new datasets, or uplift existing ones, for lesser-studied languages. More datasets for minoritised languages, which frequently document rare linguistic features (Mithun, 2007), will bring more precision to our measurements of the synchronic distribution of linguistic features. Better data availability for historical languages, especially where comparable data for cognate or daughter languages is also available, will improve our understanding of the dynamics underlying language evolution.

The Goidelic languages are an obvious target case for improved data development. They comprise a well-defined language cluster exhibiting features diverging from the areal and cross-linguistic norm at all levels of linguistic structure. All surviving Goidelic languages – Irish[3], Manx[4], and Scottish Gaelic[5] – are minoritised. While their development from Old Irish (600-900CE) is well-documented in the textual record, it has not been comprehensively described.

Old Irish itself is the earliest Celtic language for which attestation is copious enough to allow for a full grammatical description. It is noticeably divergent from related Indo-European languages. Syntactically, like other Insular Celtic languages, it has dominant verb-initial word order (Thurneysen, 1946). Morphologically, it shows extremely complex patterns of verbal inflection, even by the standards of older Indo-European languages (McCone, 1987). Phonologically, it has a large consonant system and has been described as having a vertical vowel system (Anderson, 2016). Given this linguistic profile, good computational resources for the language are an urgent desideratum.

Our contributions in this paper are the following:

[1] DOI: 10.5281/zenodo.10898227; repository: https://github.com/cormacanderson/Goidelex
[2] ISO 639-3 code `sga`; Glottocode `oldi1246`
[3] ISO 639-3 code `gle`; Glottocode `iris1253`
[4] ISO 639-3 code `glv`; Glottocode `manx1243`
[5] ISO 639-3 code `gla`; Glottocode `scot1245`

LT4HALA 2024@LREC-COLING-2024, pages 1–10
25 May, 2024.

- A lexical resource for Old Irish, interoperable with existing resources (CorPH, DIL), and providing a unified, standardised representation of lexemes and structured groupings into lexemes and flexemes (Fradin and Kerleroux, 2003; Thornton, 2018; Pellegrini, 2023).
- Normalised orthography, providing a single identifier for orthographic variants of a single lexeme.
- Generated phonological forms, facilitating morphological and phonological research.
- Detailed morphosyntactic and morphonological annotation, including part of speech, inflectional class, gender, propensity to syncope, etc.
- Information on etymology and derivational family for each lexeme.
- A manually curated set of rules for grapheme-to-phoneme conversion, starting from the normalised orthography.
- Progress towards the digitisation of the Würzburg glosses.

2. Previous work

The main digital lexical resources available for Old Irish are the electronic Dictionary of the Irish Language (DIL: Toner et al. (2013-present) and the Corpus PalaeoHibernicum (CorPH: Stifter et al. (2021)).

DIL is a longstanding dictionary resource, originally available in print format, but in recent times also online. Its lexical coverage of the language is comprehensive, but search and filter functions are quite rudimentary and the orthography of headwords inconsistent, making the assembly of examples for linguistic research very difficult. Furthermore, examples are only sometimes annotated for morphosyntactic features, making it difficult to use DIL for morphological investigation. Further, DIL has not been digitised in a way that makes it easy to extract the data computationally.

CorPH attempts to resolve these problems. While it operates over a smaller corpus than DIL, it is far more thorough in terms of morphosyntactic annotation, making it much more useful for morphological research. It still suffers from certain limitations, however, which create difficulties in terms of aggregating data. In particular, the orthography of headwords is not fully standardised, so lexemes with the same phonological profile may be spelled differently, and there are occasional duplicate entries where a single lexeme has two separate entries with differing orthography. It also does not include the Würzburg glosses, one of the most important contemporary sources for Old Irish.

At present, no digital lexical resource exists for the Würzburg glosses. While a digital edition of the text (Doyle, 2018) and a UD treebank containing a small selection of glosses are available (Doyle, 2023), these resources do not provide fine-grained phonological or morphological annotation. The most comprehensive source, and therefore most suitable for our purposes, remains the printed lexicon (Kavanagh and Wodtko, 2001).

The limitations of these existing resources create difficulties both for end-users and for linguistics researchers. For the end-user, it is difficult to find lexemes, as there is no orthographic normalisation, meaning one must try variant spellings until one finds the lexeme one is looking for. Compounding this, the search and filter capabilities in DIL are very limited, although CorPH is considerably better in this respect. For the researcher, orthographic inconsistency makes it very difficult to assemble examples for linguistic comparison and has impeded the development of standard NLP tools such as grapheme-to-phoneme conversion.

Goidelex aims to address these limitations. It provides a consistent and standardised lexical resource that will be useful as a lexical resource for both studies in Old Irish phonology, morphology and lexicon, and wider comparative linguistics research. Beyond its standalone value, it has broader function as a basis from which to produce other lexical resources, such as inflected lexicons for morphology (e.g. the Paralex datasets, Beniamine et al. 2023), concept and cognacy-coded word lists for historical linguistics (e.g. the CLDF datasets, Forkel et al. 2018), and, following Mambrini and Passarotti (2023), a lemma collection modelled as a knowledge graph according to Ontolex, the W3C de-facto standard for lexical information in the Linked Data paradigm (McCrae et al., 2017).

Goidelex focuses on the lexicon of the Würzburg glosses in the first instance, as this material is not available through CorPH. While the initial dataset has only nominal lexemes, the database will be expanded to include other parts of speech in the near future.

3. Design principles

A first problem to be confronted when developing a lexical resource for Old Irish is the ambiguous and inconsistent nature of the language's orthography. As mentioned in § 2, inconsistent spelling of headwords makes it difficult to search a resource for a given lexeme or to filter lexemes to draw up a list of examples for research. This inconsistency also hampers the development of NLP tools, such

as grapheme-to-phoneme conversion. Our solution to this problem in Goidelex was to use a normalised orthography (see § 3.1).

Beyond orthography, the Old Irish lexicon exhibits considerable variation at all levels of linguistic structure. In some cases, the same lexeme shows different phonological forms across surviving corpora. In others, there are differences in inflection, be it in morphonological behaviour, such as the occurrence of syncope in certain forms, or in morphosyntactic properties such as gender or inflectional class. We attempt to capture this variation by a principled distinction between lexemes and flexemes (§ 3.2).

3.1. Normalised orthography

A key innovation of Goidelex is the use of a normalised orthography for citation forms. This has a number of advantages. First, existing sources frequently differ in the spelling of the citation forms they use for any given lexeme, which makes it difficult to identify lexemes within and across sources. The normalised orthography provides a principled representation that makes it easier for users of the database to find lexemes. Second, it provides a human readable form that serves as an identifier to link data from different corpora. As such, it is a secure basis for lemmatisation, reducing considerably the risk of duplicate headwords (as occur occasionally in Stifter et al., 2021). Third, and most critically, it constitutes a standardised starting point for grapheme-to-phoneme conversion (§ 4.3).

We follow the normalised orthography proposed by Fransen et al. (2023), which adheres to six basic principles: comprehensiveness, clarity, neutrality, redundancy, fidelity, and conventionality. It is intended to represent all possible forms in Old Irish. Each normalised orthographic form aims to correspond to a single phonological form, while for each phonological form there is a single, obvious, orthographic representation. The normalised orthography remains as neutral as possible with respect to different phonological analyses and makes ample use of redundancy in cases of uncertainty. It aims to be as faithful as possible to genuine Old Irish spelling and to existing scholarly conventions.

3.2. Lexemes and flexemes

In Goidelex, we take lexemes to be defined by a shared meaning and a single part of speech. This means, for example, that deadjectival nouns are to be listed separately from the adjectives from which they derive, and denominal verbs are to be listed separately from the nouns from which they are formed. Derivational relationships between lexemes are captured by the notion of derivational families (§ 5.2).

However, a single lexeme sometimes leads to multiple distinct inflectional paradigms, due to variation in its phonology, morphonology, or morphosyntactic behaviour. To capture this variation, we use the notion of *flexeme* (Fradin and Kerleroux, 2003; Thornton, 2018; Pellegrini, 2023). In this approach, each inflectional variant of a lexeme, differing in terms of phonology, morphonology, or morphosyntax, is analysed as a separate flexeme. Thus, a single lexeme may map to multiple flexemes.

Some examples can serve to illustrate this. The noun *muinter* 'family, household' sometimes appears as *muinter* and sometimes as *muntar*. These different spellings reflect a phonological difference: the cluster is palatalised /nʲtʲ/ in the first instance and labiovelarised /nʷtʷ/ in the second. On this basis, we have two separate flexemes, both linked to the same lexeme entry. A futher example is provided by the noun *fius*, which does not vary in terms of its phonology, but which varies with respect to morphosyntactic category. Sometimes it is inflected as a neuter u-stem, sometimes as a masculine u-stem, and sometimes as a neuter o-stem. We thus set up three different flexemes corresponding to this single lexeme.

Identifying flexemes required detailed manual study of the attested forms of each lexeme appearing in the Würzburg and CorPH datasets. A total of 107 out of 574 lexemes showed variation either in terms of their phonology, their morphonological patterning, their gender, or their inflectional class. In total, there are 695 flexemes in the Goidelex launch dataset, corresponding to 574 lexemes.

4. Building the database

We produced the database in three steps. First, we manually entered lemmata from the Würzburg glosses into the `Lexeme` and `Flexeme` tables (§ 4.1). Then, we merged lexemes with CorPH lemmata in a semi-automatic fashion (§ 4.2). Finally, we carried out automatic grapheme-to-phoneme conversion using customised rules (§ 4.3). Further tables were input manually.

4.1. Würzburg lemmata entry

The first stage of data collection involved manually entering nouns from the Würzburg glosses. This corpus was chosen as it is by far the largest and most important corpus of Old Irish for which no digital lexical resource was available. All nouns with more than one attestation in the lexicon of the Würzburg glosses (Kavanagh and Wodtko, 2001) were included, amounting initially to a total of 574

nouns.

This yielded a list of orthographic headwords, to which we manually added a detailed part of speech, gender, inflection class, gloss, derivational family annotation, and url references to entries in the electronic Dictionary of the Irish language (Toner et al., 2013-present). Detailed study of the Würzburg glosses was necessary in order to identify orthographic, morphological or phonological variation and conduct a preliminary analysis of entries into lexemes and flexemes. To facilitate bridging across resources as well as phonological transcription, we manually transcribed each lemma into the normalised orthography proposed by Fransen et al. (2023).

4.2. Merging with CorPH lemmata

Lexical entries were then aligned with corresponding data in CorPH (Stifter et al., 2021). First, we manually annotated each lexeme with the corresponding headwords in CorPH. Then, leveraging these headwords, we automatically extracted from CorPH lemma ID numbers, full meaning definitions (more complete than our short glosses), and etymological information. We flagged potential problems and manually corrected all cases in which there were mismatches between our annotations and those found in CorPH. In certain instances, this involved adding also new flexemes to capture variation in the CorPH dataset that is not present in the Würzburg corpus.

4.3. Grapheme-to-phoneme conversion

We then generated phonological forms from the lexemes in normalised orthography using handmade rules. As well as being suitable for cross-linguistic comparison, phonological forms are a principled basis from which to develop new tools and resources for Old Irish.

We write phonological forms according to the phonological system set out in Anderson (2016). In order to convert normalised orthographic forms to this representation we used the Epitran software (Mortensen et al., 2018) to process our grapheme-to-phoneme rules. Epitran proceeds in three steps, each applied independently on input forms (here citation forms in normalised orthography):

1. **Preprocessing**: a first set of ordered rules.
2. **Mapping**: a set of non-contextual mappings.
3. **Postprocessing**: a second set of ordered rules.

Epitran rules are written according to a custom syntax that resembles traditional phonological rules, employing variables and regular expressions. We devised our own set of rules and grouped them into numbered blocks to facilitate readability, identification of errors, and validation.

5. Structure of the database

The database is structured as a set of CSV files, linked by foreign key relations (see Figure 1). Since no standard existed for the specific type of data in question here, we chose formats and structures compatible with related standards. In particular, Goidelex is designed to be easily extended (see § 6) into datasets fitting either the Paralex standard for inflected lexicons (Beniamine et al., 2023) or the Cross-Linguistic Data Format standard suitable for cognate-coded lexical data (Forkel et al., 2018).

5.1. Lexemes table

The `Lexemes` table (Table 1.a) identifies individual lexemes and links these to other Old Irish resources. Lexemes are identified by a unique identifier, as well as by a human readable citation form written in the normalised orthography developed for this project. Entries in the `Lexemes` table are linked to two online Old Irish resources (Stifter et al., 2021; Toner et al., 2013-present) and to the Würzburg dictionary (Kavanagh and Wodtko, 2001).

Information about lexemes was mostly manually annotated by the authors, but in some cases was drawn from CorPH. Each lexeme is defined as belonging to a single part of speech, meaning, for example, that adjectives have separate lexical entries to deadjectival nouns formed from them, while verbal nouns are listed separately to the verbs to which they are associated. However, related lexemes are aggregated into derivational families (§ 5.2).

- **lexeme_id**: The primary key for this table, it identifies the entire row and acts as a foreign key in the `Flexemes` table (§ 5.3).
- **derivational_families**: Foreign key identifier(s) from the `Derivational families` table (§ 5.2), separated by semicolons where more than one entry.
- **label**: Human readable citation form for the lexeme in normalised orthography.
- **CorPH_ids**: The identification number(s) of the corresponding lexeme in the CorPH database.
- **CorPH_labels**: The citation form(s) of the corresponding lexeme in the CorPH database.

(a) Lexemes table

lexeme_id	derivational _families	label	CorPH_ids	CorPH_labels	CorPH_meaning	Wb_label	DIL_URL	gloss	POS
lex-apstal-56	55	apstal	3389	apstal	apostle	apstal, abstal	http://dil.ie/3887	apostle	noun
lex-apstalach-57	55	apstalach	3390	apstalacht	apostleship, apostolate	apstalacht	http://dil.ie/3889	apostolate	noun
lex-brithem-98	92	brithem	3573	brithem	judge	brithem	http://dil.ie/6699	judge	noun
lex-firinne-280	262	fírinne	4620	fírinne	truth; justice, righ[...]	fírinne	http://dil.ie/22203	righteousness, truth	noun
lex-fius-282	54	fius	4627	fius	the act of finding o[...]	fiuss, fius, fis	http://dil.ie/22221	knowledge	verbal_noun
lex-muinter-429	400	muinter	2187	muinter	family, household, f[...]	muntar	http://dil.ie/32754	community	noun
lex-talam-525	482	talam	5880	talam	earth, world; ground[...]	talam	http://dil.ie/39932	earth	noun

(b) Flexemes table

flexeme_id	label	lexeme	texts	etymology	inherent_properties	phonological_form
apstal-56	apstal	lex-apstal-56			sync_none;alt_none;gen_masc;stem_o;num_all	ˈØˠapˠsˠtˠalˠ
apstalach-57	apstalach	lex-apstalach-57			sync_none;alt_none;gen_fem;stem_ā;num_all	ˈØˠapˠsˠtˠalˠaxˠ
brithem-98	brithem	lex-brithem-98	79;6;12	denominative (+ agent suffix -em): < breth	sync_none;alt_none;gen_masc;stem_n;num_all	ˈbʲrʲəθʲaμˠ
breithem-98.1	breithem	lex-brithem-98	5	denominative (+ agent suffix -em): < breth	sync_none;alt_none;gen_masc;stem_n;num_all	ˈbʲrʲaθʲaμˠ
firinne-280	fírinne	lex-firinne-280		denominative (abstract): < fírión/fírián/fíré[...]	sync_vf;alt_none;gen_fem;stem_iā;num_all	ˈɸʲəØʲrʲəɴʲaØʲ
fius-282	fius	lex-fius-282		*u̯id-tu-; vn. of ro·fitir	sync_none;alt_none;gen_neut;stem_u;num_all	ˈɸʲəsʷ
fius-282.1	fius	lex-fius-282		*u̯id-tu-; vn. of ro·fitir	sync_none;alt_none;gen_masc;stem_u;num_all	ˈɸʲəsʷ
fius-282.2	fius	lex-fius-282		*u̯id-tu-; vn. of ro·fitir	sync_none;alt_none;gen_neut;stem_o;num_all	ˈɸʲəsʷ
muinter-429	muinter	lex-muinter-429	79;70;5;1	<Lat. monasterium? via British?	sync_none;alt_none;gen_fem;stem_ā;num_all	ˈmʷənʲtʲarˠ
muntar-429.1	muntar	lex-muinter-429	6;7	<Lat. monasterium? via British?	sync_none;alt_none;gen_fem;stem_ā;num_all	ˈmʷənʷtʷarˠ
talam-525	talam	lex-talam-525		*telamon-	sync;alt_none;gen_masc;stem_n;num_all	ˈtˠalˠaμˠ

(c) Inherent properties table

properties_id	label	comment	domain	type
alt_none	No morphonological alternations	This flexeme does not show morphonological alternations	morphonology	alternation
sync_none	No syncope	The vowel of the final syllable of this flexeme is not liable to syncope	morphonology	syncope
sync_vf	Vowel final	This flexeme ends in a vowel, which is liable to be lost	morphonology	syncope
sync	Syncope	The vowel of the final syllable of this flexeme is liable to syncope	morphonology	syncope
gen_masc	Masculine	This flexeme has masculine gender	morphosyntax	gender
gen_fem	Feminine	This flexeme has feminine gender	morphosyntax	gender
gen_neut	Neuter	This flexeme has neuter gender	morphosyntax	gender
stem_o	o-stem noun	This flexeme is inflected as an o-stem	morphosyntax	nominal_stem
stem_ā	ā-stem noun	This flexeme is inflected as an ā-stem	morphosyntax	nominal_stem
stem_n	n-stem noun	This flexeme is inflected as an n-stem	morphosyntax	nominal_stem
stem_iā	iā-stem	This flexeme is inflected as an iā-stem	morphosyntax	nominal_stem
stem_u	u-stem noun	This flexeme is inflected as a u-stem	morphosyntax	nominal_stem
num_all	No number restriction	This flexeme is inflected for all numbers	morphosyntax	number_restriction

Table 1: Excerpts from the Lexeme, Flexeme and Inherent Properties tables (long cells are truncated).

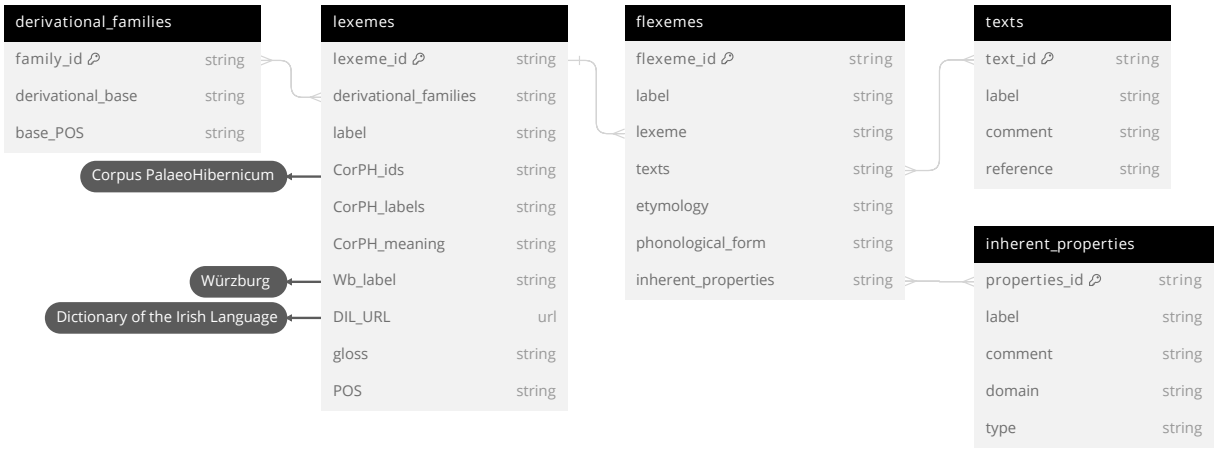

Figure 1: Database schema. External resources are indicated as rounded boxes.

- **CorPH_meaning**: A full meaning description for the lexeme extracted from the CorPH database.
- **DIL_url**: The url of the corresponding lexeme in the online Dictionary of the Irish Language (DIL).
- **Wb_label**: The citation form of the corresponding lexeme in the lexicon of the Würzburg glosses (Kavanagh and Wodtko, 2001).
- **gloss**: A short description of the meaning of the lexeme.
- **POS**: The part of speech of the lexeme.

5.2. Derivational families table

family_id	derivational_base	base_POS
55	apstal	noun
92	beirid	verb
262	fíor	adjective
54	ro·fitir	verb
400	muinter	noun
482	talam	noun

Table 2: A few rows from the Derivational families table

The `Derivational families` table (Table 2) links together lexemes that stand in a derivational relationship (including compounds). In some cases, this does not involve a change in part of speech, e.g. *apstal* 'apostle' and *apstalacht* 'apostolate' are both nouns. In other cases, linked lexemes can have different parts of speech, e.g. *fíor* 'true' is an adjective, whereas *fírinne* 'truth' is a noun.

- **family_id**: The primary key for this table, it identifies the entire row and acts as a foreign key in the `Lexemes` table (§ 5.1).
- **derivational_base**: The citation form in normalised orthography for the derivational base of a lexeme.[6]
- **base_POS**: The part of speech of the derivational base.

5.3. Flexemes table

A single lexeme can sometimes show phonological, morphonological or morphosyntactic variation. We thus define flexemes as finely-grained variants of lexemes (Fradin and Kerleroux, 2003; Thornton, 2018; Pellegrini, 2023), each belonging to a single inflectional microclass (Dressler, 2002; Beniamine et al., 2018). In Goidelex, each flexeme has a unique identifier and a label in normalised orthography (Table 1.b). It is linked by foreign keys to its parent lexeme and we provide further information regarding textual distribution, inflection class, etymology, and any morphonological particularities that are not predictable from its orthographic form.

- **flexeme_id**: The primary key for this table, this identifies the entire row. Derived resources are expected to refer to these identifiers.
- **label**: A human readable label for the flexeme in normalised orthography.
- **lexeme**: A foreign key identifying the parent lexeme in the `Lexemes` table (§ 5.1).

[6] At this point, we treat the citation form of the associated verb as the derivational base of a verbal noun. This idealisation will facilitate expansion of the dataset to also include verbal forms.

- **texts**: The set of available texts in which this variant occurs, given as text codes separated by a semicolon. Text codes are foreign key identifiers from the `Texts` table (§ 5.5).

- **etymology**: The etymology of the flexeme, drawn from CorPH.

- **inherent_properties**: A set of foreign keys, separated by semicolon, identifying non-predictable morphonological or morphosyntactic information about the lexeme in the `Inherent properties` table (§ 5.4).

- **phonological_form**: A phonological form generated by grapheme-to-phoneme conversion.

Full normalisation of `inherent_properties` would have required an intermediate table mapping identifiers from the flexeme and inherent properties tables. However, such a (very long) table would be nearly unusable to users reading the database in spreadsheet software, or who may not be able to perform database joins. Conversely, setting the properties in wide format (as is often the preference for qualitative research), with columns for each type of property, would have made the table very specific to nominal entries, and would lead to a large increase in columns for verbal entries. Our choice here is instead a compromise between normalisation and ease of use: by keeping the long form, we can fully describe each property in the relevant table (§ 5.4), while ensuring that these properties can be read directly from the relevant rows of the `Flexemes` table, which is more intuitive to less technical users. This comes at the cost of adding cell-internal separators (here, semicolons), a choice we resorted to also for a few other columns, such as `flexemes.texts`, `lexemes.CorPH_ids`, `lexemes.CorPH_labels`.

5.4. Inherent properties table

As described in § 3.2, some flexemes have inherent properties (Table 1.c) that are not predictable from their normalised orthographic form. These include morphonological properties as well as inherent morphosyntactic information such as gender and inflection class.

A common example of a non-predictable morphonological property of a flexeme, which is annotated in this table, is propensity to syncope. For example, *talam* 'land' and *brithem* 'judge' are both masculine n-stem nouns. However, *talam* has the genitive singular form *talman*, with syncope of the second syllable, while *brithem* has the genitive singular form *britheman*, without syncope.

Morphonosytactic properties annotated here include gender (masculine, neuter, and feminine in Old Irish), and inflectional class. As with the morphonological properties, these morphosyntactic properties are not predictable from the orthographic form, so must be annotated for each individual flexeme.

The `Inherent properties` table lists all valid codes for these properties. The launch version of Goidelex has only nouns, so currently only properties relevant to nouns need to be annotated. Further rows will be added here in future as we expand the database to include also other parts of speech.

- **properties_id**: The primary key of this table and a foreign key in the `Flexemes` table (§ 5.3).

- **label**: A human readable label identifying the phonological or morphonological property to which the identifier refers.

- **comment**: The text description of the property described.

- **domain**: Properties pertain to different linguistic domains. The domains in use are `morphonology` and `morphosyntax`.

- **type**: Properties can be logically grouped. This field assigns a type grouping to each class identified. The types in use are: `gender`, `stem_class`, `alternation`, `syncope`, `number_restriction`.

5.5. Texts table

Some flexemes occur in one set of texts, while others occur in other texts within the Old Irish corpus. This table (Table 3) provides explicit information regarding the texts in which a particular variant is found.

- **text_id**: The primary key of this table and a foreign key in the `Flexemes` table (§ 5.3). The text IDs are the same as those used in CorPH, with some extension to include texts (predominantly the Würzburg glosses) which do not occur in that dataset.

- **label**: A human readable label identifying the text.

- **comment**: A text description of the text in question.

- **reference**: A bibliographic reference for this text, in most cases also taken from CorPH.

text_id	label	Comment	reference
1	Annals of Ulster		Mac Airt and Mac Niocaill 1983
2	Vita Columbae		Anderson and Anderson 1961; Thes. II, 272–280
3	Baile Chuinn		Murray and Bhreathnach 2005
4	Disciples and Relatives of Columba		Anderson and Anderson 1961;Thes. II, 281
5	Poems of Blathmac		Barrett 2018; Carney 1964

Table 3: A few rows from the texts table

6. Conclusion and future work

Goidelex has been designed to act as a central lexical resource for Old Irish. It aligns data from multiple sources, provides central identifiers and normalised representations, as well as very detailed phonological and morphological annotation. The database makes this information accessible for a wide range of qualitative and quantitative purposes.

Many open questions about Old Irish phonology and morphology can be addressed using the database. The structured nature of the Goidelex data makes it easy to collect examples for investigation from corpora, something which is difficult or impossible with existing resources. Consistent annotation of phonological and morphosyntactic properties opens up numerous possibilities for research into Old Irish phonology and morphology, while the grouping of lexemes into derivational families facilitates studies of word formation.

As a central resource, Goidelex lends itself to extensions as separate datasets. In particular, we envision three types of derived datasets: inflected lexicons, cross-linguistic cognacy datasets, and a linked lemma bank.

Goidelex constitutes a sound basis from which to develop an inflected lexicon of the Old Irish noun, compatible with the Paralex standard (Beniamine et al., 2023). Indeed, the surface inflectional paradigms of each flexeme are fully predictable from the phonological transcriptions and the morphonological and morphosyntactic annotations documented in Goidelex. This fine-grained information can serve as the input for finite-state transducers in order to generate full inflected paradigms. Currently, no such resource exists (the Old Irish Unimorph dataset (Batsuren et al., 2022) counts 50 verbs, only in orthography, and with no nominal paradigms).

Goidelex is also meant to serve as a basis for developing cross-linguistic comparative cognacy data for the Goidelic languages. There already exists a bridge (Scannell, 2018), which links entries between the the most important dictionaries of Old Irish (DIL: Toner et al., 2013-present) and Modern Irish (Dónaill, 1977) and similar work is under way to link Modern Irish to other modern Goidelic languages. The design of Goidelex, being compatible with the CLDF standard (Forkel et al., 2018), facilitates efforts to align cognate data to other languages.

Finally, ongoing work (Fransen et al., 2024), inspired by similar efforts for Latin (Mambrini and Passarotti, 2023), uses a subset of Goidelex to create a lemma bank for Old Irish within the Linked Data paradigm.

7. Ethical statement

To the best of our knowledge there are no ethical concerns pertaining to this resource.

8. Acknowledgements

Cormac Anderson is funded by a British Academy Grant (GP GP300169) while Sacha Beniamine is funded by a Leverhulme Early Career Fellowship (ECF-2022-286). Theodorus Fransen has received funding from the European Union's Horizon Europe scientific research initiative under the Marie Skłodowska-Curie Actions (MSCA), grant agreement No 101106220 (MOLOR – Morphologically Linked Old Irish Resource).

9. Bibliographical references

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

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David

Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

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.

Sacha Beniamine, Olivier Bonami, and Benoît Sagot. 2018. Inferring inflection classes with description length. *Journal of Language Modelling*, 5(3):465–525.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 7817–7829, Dublin. Association for Computational Linguistics.

Östen Dahl. 2015. How WEIRD are WALS languages? Paper presented at the Diversity linguistics - retrospect and prospect conference, Max Planck Institute for Evolutionary Anthropology, May 1-3, 2015, Leipzig.

Wolfgang Dressler. 2002. Latin inflection classes. In A. Machtelt Bolkestein, Caroline H.M. Kroon, Harm Pinkster, and Rodie Risselada H. Wim Remmelink, editors, *Theory and description in Latin linguistics: Selected Papers from the XIth International Colloquium on Latin Linguistics*, pages 91–110. Brill, Leiden.

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).

Bernard Fradin and Françoise Kerleroux. 2003. Troubles with lexemes. In Geert Booij, Janet DeCesaris, Angela Ralli, and Sergio Scalise, editors, *Selected papers from the third Mediterranean Morphology Meeting*, pages 177–196. IULA – Universitat Pompeu Fabra.

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. Paper accepted to the *9th Workshop on Linked Data in Linguistics (LDL-2024)* at LREC-Coling 2024.

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.

R. Malouf, F. Ackerman, and A. Semenuks. 2020. Lexical databases for computational analyses: A linguistic perspective. *Society for Computation in Linguistics*, 3(1):297–307.

Francesco Mambrini and Marco Carlo Passarotti. 2023. The LiLa Lemma Bank: A Knowledge Base of Latin canonical forms. *Journal of Open Humanities Data*, 9(28):1–5.

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

John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The OntoLex-Lemon Model: Development and Applications. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, pages 587–597, Brno, Czech Republic. Lexical Computing CZ s.r.o.

Marianne Mithun. 2007. Linguistics in the face of language endangerment. In Leo W. Wetzels, editor, *Language endangerment and endangered languages: Linguistic and anthropological studies with special emphasis on the languages and cultures of the Andean-Amazonian border area (Indigenous Languages of Latin America (ILLA)*, volume 5 of *Publications of the Research School of Asian, African, and Amerindian Studies (CNWS) 154)*, pages 15–35. Research School CNWS, Leiden University, Leiden.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Matteo Pellegrini. 2023. Flexemes in theory and in practice. *Morphology*, 33:361–395.

Anna M. Thornton. 2018. Troubles with flexemes. In Oliver Bonami, Gilles Boyé, Hélène Firaudo, and Fiammetta Namer, editors, *The lexeme in descriptive and theoretical morphology*, pages 202–321. Language Science Press, Berlin.

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

10. Language resource references

Adrian Doyle. 2018. *Würzburg Irish Glosses*. online. PID https://wuerzburg.ie/.

Adrian Doyle. 2023. *Diplomatic Würzburg Glosses Treebank (DipWBG)*. Universal Dependencies 2.13. PID https://universaldependencies.org/treebanks/sga_dipwbg/.

Niall Ó Dónaill. 1977. *Foclóir Gaeilge-Béarla*. An Gúm. PID https://www.teanglann.ie/.

Kevin Scannell. 2018. *Droichead DIL*. Online. PID https://cadhan.com/droichead/. Presentation "Is iomaí cor i saol an fhocail: Linking online dictionaries of Old Irish and Modern Irish", NAACLT confrence, St. Louis, 2018.

David Stifter and Bernhard Bauer and Elliott Lash and Fangzhe Qiu and Nora White and Siobhán Barrett and Aaron Griffith and Romanas Bulatovas and Francesco Felici and Ellen Ganly and Truc Ha Nguyen and Lars Nooij. 2021. *Corpus PalaeoHibernicum*. Maynooth University, 1.0. PID http://chronhib.maynoothuniversity.ie.

Gregory Toner and Maxim Fomin and Grigory Bondarenko and Thomas Torma and Caoimhín Ó Dónaill and Hilary Lavelle. 2013-present. *An Electronic Dictionary of the Irish Language*. Royal Irish Academy. PID https://www.dil.ie. Based on the Contributions to a Dictionary of the Irish Language, 1913-1976.