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

Theodorus Fransen, Federico Simone Samperi, Elisa Roma, Paolo Ruffolo, and Marco Passarotti

LIFT2 | Orléans La Source campus, France | 14–15 November 2024

UNIVERSITÀ CATTOLICA del Sacro Cuore

CIRCSE
Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione

Funded by the European Union

This project has received funding from the European Union's Horizon Europe scientific research initiative under the Marie Skłodowska-Curie Actions (MSCA) - grant agreement no. 101106220

Overview

- Project overview: MOLOR
- Old Irish: linguistic and resource context
- Linguistic Linked Open Data, LiLa & OntoLex
- Populating the MOLOR Lemma Bank: alignment procedures
- Conclusions and future work

MOLOR: aims and objectives

MOLOR

Aims and objectives

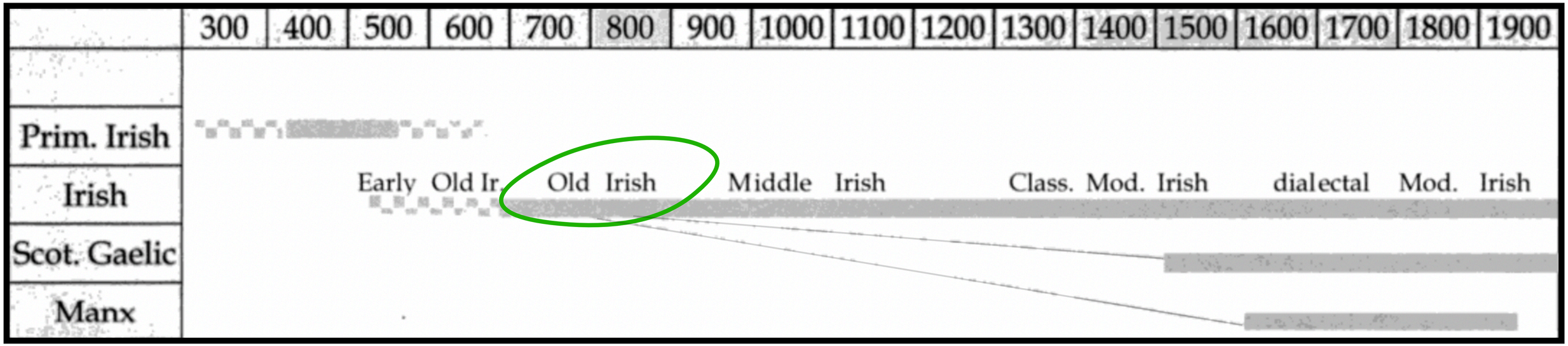
- Morphologically Linked Old Irish Resource
- A 2-year Marie Skłodowska-Curie postdoctoral fellowship project at the intersection of lexicography, morphology and the semantic web for Old Irish
- Current focus: a Lemma Bank as a hub in a Linguistic Linked Open Data architecture for Old Irish
- Expected Impact:
  - Scientific: New standards for historical language lexicography (at least for Old Irish)
  - Educational: Better tools for teaching Old Irish
  - Technological: Usage and refinement of linguistic linked data models and standards
  - Cultural: Preservation and accessibility of European linguistic heritage

Old Irish: linguistic context

Old Irish

Goidelic languages: dating

- Indo-European > Celtic > Insular Celtic > **Goidelic**

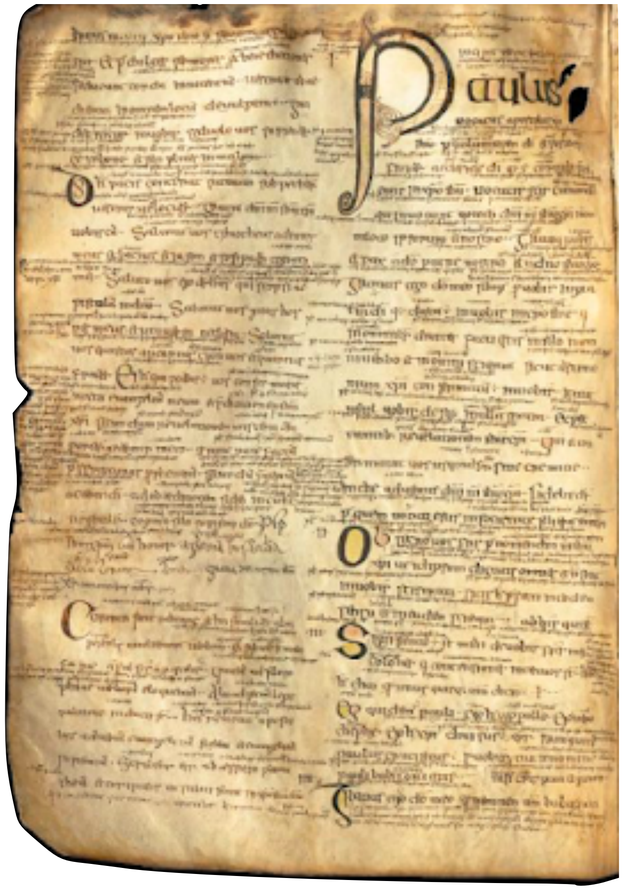


Source: Stifter, David. 2006. *Sengoídelc. Old Irish for Beginners*. Syracuse University Press, p. 7.

Old Irish

Linguistic background[2]

- Mostly glosses (in Irish and Latin), most of which from manuscripts c. 700–850CE
- Old Irish is characterised by morpho(phono)logical complexity and orthographic variability — otherwise fairly homogeneous
- Middle Irish (c. 900–1200CE): unstable phase showing grammatical simplification/typological changes and much linguistic variation
- Old Irish texts survive in manuscripts from the Middle Irish period -> Old+Middle Irish = “Early Irish”



Würzburg glosses, (~800CE), folio 7v.
Source: http://anglandicus.blogspot.com/2011/03/wurzburg-glosses.html

2. Stifter, David. 2009. Early Irish. In Martin J. Ball & Nicole Müller (eds.). *The Celtic languages*, 2nd edn., 55–116. Abingdon & New York: Routledge

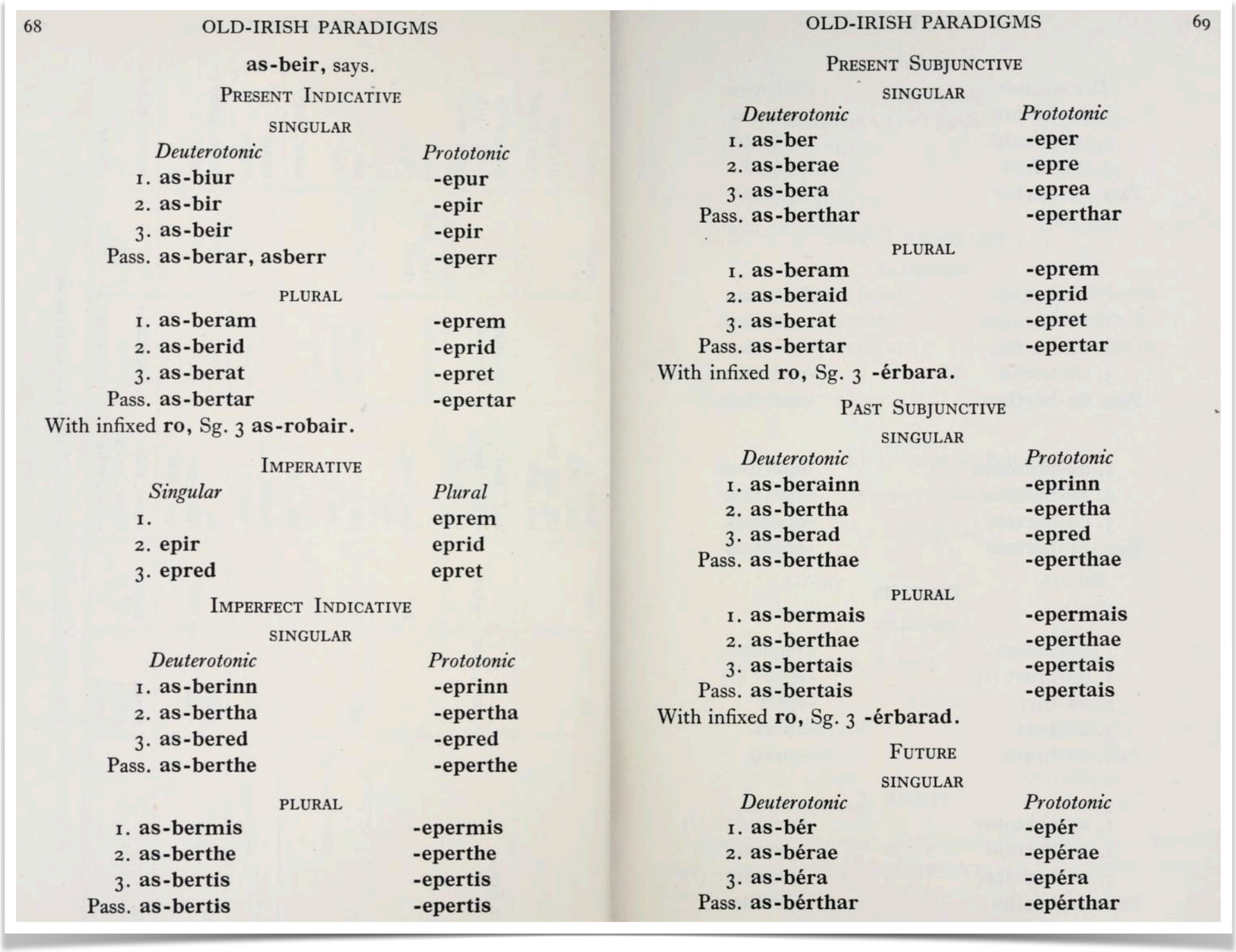
Old Irish

A selection of (challenging) grammatical features

- Orthography
  - Variability & opaque grapheme-phoneme relationship, e.g. *corp* ‘body’ (with /p/), *carpat* ‘chariot’ (with /b/), *burb(b)ae*, *burpe* ‘stupidity’ (also with /b/) but *murbach, muirbech* ‘breakwater’ (with /v/)
  - Segmentation/tokenisation, e.g (*hore*) *nondobmolorsa* ‘(because) I praise ye’ (Würzburg 14c18)
- Phonology
  - Initial mutations, e.g. *catt* (/kat/) ’cat’ ~ *in chaitt* (/xatʲ/) ‘the cat’s’ (gen.sg)
  - Grammaticalised palatalisation (and velarisation) — see above example
  - Ablaut, e.g. *fer* ‘man’ ~ *fir* (gen.sg), *fiur* (dat.sg)
- Morphology
  - Intricate allomorphy, most notoriously in the verbal system (next slide), e.g. *as·beir* ~ *·epir* ‘says’ (*·epeir*, *·eiper*, *·eper*)
- (Morpho)syntax
  - VSO
  - Relativity: no relative pronouns in the language, use of special verbal endings or initial mutations instead

Old Irish

Double stem formation with verbs 😱



68 OLD-IRISH PARADIGMS

as-beir, says.

PRESENT INDICATIVE

SINGULAR

	Deuterotonic	*Prototonic*
1.	as-biur	-epur
2.	as-bir	-epir
3.	as-beir	-epir
Pass.	as-berar, asberr	-eperr

PLURAL

1.	as-beram	-eprem
2.	as-berid	-eprid
3.	as-berat	-epret
Pass.	as-bertar	-epertar

With infixed **ro**, Sg. 3 **as-robair**.

IMPERATIVE

	Singular	*Plural*
1.		eprem
2.	epir	eprid
3.	epred	epret

IMPERFECT INDICATIVE

SINGULAR

	Deuterotonic	*Prototonic*
1.	as-berinn	-eprinn
2.	as-bertha	-epertha
3.	as-bered	-epred
Pass.	as-berthe	-eperthe

PLURAL

1.	as-bermis	-epermis
2.	as-berthe	-eperthe
3.	as-bertis	-epertis
Pass.	as-bertis	-epertis

OLD-IRISH PARADIGMS 69

PRESENT SUBJUNCTIVE

SINGULAR

	Deuterotonic	*Prototonic*
1.	as-ber	-eper
2.	as-berae	-epre
3.	as-bera	-eprea
Pass.	as-berthar	-eperthar

PLURAL

1.	as-beram	-eprem
2.	as-beraid	-eprid
3.	as-berat	-epret
Pass.	as-bertar	-epertar

With infixed **ro**, Sg. 3 **-érbara**.

PAST SUBJUNCTIVE

SINGULAR

	Deuterotonic	*Prototonic*
1.	as-berainn	-eprinn
2.	as-bertha	-epertha
3.	as-berad	-epred
Pass.	as-berthae	-eperthae

PLURAL

1.	as-bermais	-epermais
2.	as-berthae	-eperthae
3.	as-bertais	-epertais
Pass.	as-bertais	-epertais

With infixed **ro**, Sg. 3 **-érbarad**.

FUTURE

SINGULAR

	Deuterotonic	*Prototonic*
1.	as-bér	-epér
2.	as-bérae	-epérae
3.	as-béra	-epéra
Pass.	as-bérthar	-epérthar

Source: Strachan, John. 1949. *Old-Irish paradigms, and selections from the Old-Irish glosses*. 4th edn. Originally published in 1904/1905. Revised by Osborn Bergin, with notes and vocabulary. Dublin: Royal Irish Academy

Old Irish

Double stem formation with verbs (cont'd)

PLURAL

	Deuterotonic	*Prototonic*
1.	as-béram	-epéram
2.	as-béraid	-epéraid
3.	as-bérat	-epérat
Pass.	as-bértar	-epértar

SECONDARY FUTURE

SINGULAR

	Deuterotonic	*Prototonic*
1.	as-bérainn	-epérainn
2.	as-bértha	-epértha
3.	as-bérad	-epérad
Pass.	as-bérthae	-epérthae

PLURAL

1.	as-bérmais	-epérmais
2.	as-bérthae	-epérthae
3.	as-bértais	-epértais
Pass.	as-bértais	-epértais

PRETERITE

SINGULAR

	Deuterotonic	*Prototonic*
1.	as-biurt (?)	-epurt (?)
2.	as-birt	-epirt
3.	as-bert	-epert
Pass.	as-breth	-epred

PLURAL

1.	as-bertmar	-epertmar (?)
2.	as-bertaid	-epertaid (?)
3.	as-bertatar	-epertatar (?)
Pass.	as-bretha	-epertha

PERFECT

SINGULAR

	Deuterotonic	*Prototonic*
1.	as-ruburt	-érburt
2.	as-rubairt	-érbairt
3.	as-rubart	-érbart
Pass.	as-robrad	-érbrad

PLURAL

1.	as-rubartmar	-érbartmar
2.	as-rubartaid	-érbartaid
3.	as-rubartatar	-érbartatar

Verbal of Necessity : **eperthi.**

Verbal Noun : **epert.**

do-beir, gives, brings.

PRESENT INDICATIVE

SINGULAR

	Deuterotonic	*Prototonic*
1.	do-biur	-tabur
2.	do-bir	-tabair
3.	do-beir	-tabair
Pass.	do-berar, doberr	-tabarr

PLURAL

1.	do-beram	-taibrem
2.	do-berid	-taibrid
3.	do-berat	-taibret
Pass.	do-bertar	-tabartar

Perfective Present corresponding to **do-ratus,** Sg. 3 **do-rati, -tarti;** corresponding to **do-uccus,** Sg. 3 **do-uccai, -tuccai;** Pass. Sg. **-tucthar.**

Old Irish: resource context

Old Irish

Fragmented resources

- Isolated resources and projects: dictionaries and lexical databases (Griffith et al., 2018),[3] corpora, the beginnings of UD treebanks, …
- Würzburg glosses until very recently not available in machine-readable/linguistically annotated format — but see Doyle (2018; 2023)[4,5] and Anderson et al. (2024)[6]
- Previous text-and-lexicon interlinking projects (Nyhan 2006; 2008)[7,8] mostly obsolete/dormant
  - Wordlink (Ó Donnaíle, 2014)[9] connects arbitrary webpages to dictionary entries but does not generally perform well on Old Irish inflected forms and is not built on Linked Data principles

3. Griffith, A., Stifter, D., & Toner, G. 2018. Early Irish lexicography - A research survey. *Kratylos*, 63(1). 1–28.
4. Doyle, Adrian, Würzburg Irish Glosses (2018), <www.wuerzburg.ie> [accessed 9 September 2024]
5. Doyle, Adrian. 2023. Diplomatic Würzburg Glosses Treebank. Universal Dependencies. <https://universaldependencies.org/treebanks/sga_dipwbg/index.html> [accessed 9 September, 2024]
6. Anderson, Cormac, Sacha Beniamine, & 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*, 1–10, Torino, Italia. ELRA and ICCL
7. Nyhan, Julianne. 2006a. *The application of XML to the Historical Lexicography of Old, Middle and Early Modern Irish: a Lexicon-based Analysis*. Ph.D. thesis. University College Cork
8. Nyhan, Julianne. 2008. Developing integrated editions of minority language dictionaries: the Irish example. *Literary and Linguistic Computing* 23(1). 3–12.
9. 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 (eds). *Proceedings of the First Celtic Language Technology Workshop*, 18–27, Dublin, Ireland: Association for Computational Linguistics and Dublin City University

Linguistic investigation is hindered by scattered and unconnected lexical and textual resources

Linguistic Linked Open Data, LiLa & OntoLex

Linked Data

Design principles by Tim Berners-Lee (https://www.w3.org/DesignIssues/LinkedData.html)

Linked Data

make Internet data machine-readable and interconnected

The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.

Like the web of hypertext, the web of data is constructed with documents on the web. However, unlike the web of hypertext, where links are relationships anchors in hypertext documents written in HTML, for data they links between arbitrary things described by RDF,. The URIs identify any kind of object or concept. But for HTML or RDF, the same expectations apply to make the web grow:

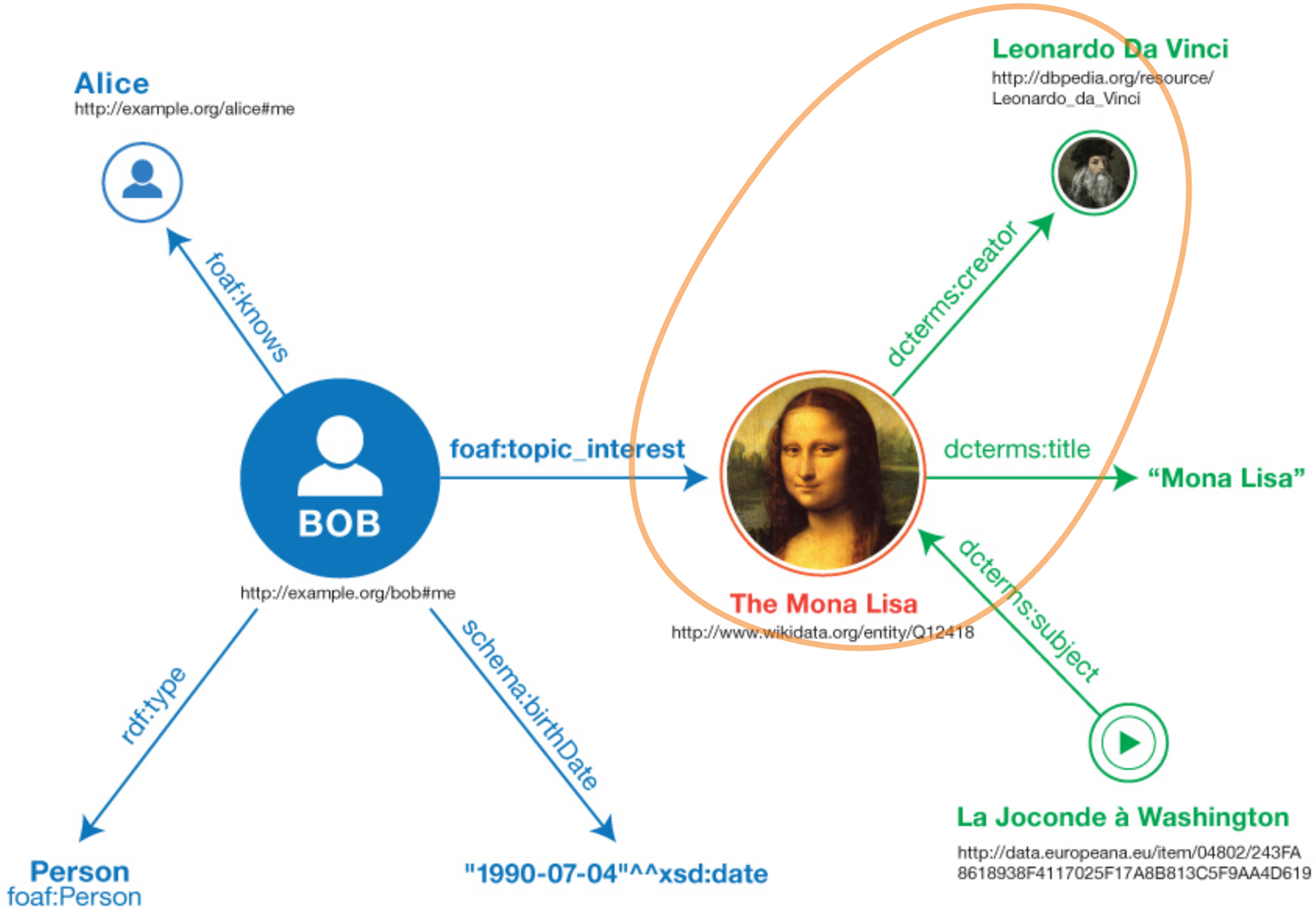
e.g. entry in a lexicon, token in a corpus

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs. so that they can discover more things.

W3C standards based on flexible and expressive graph-based data models

Knowledge Graphs

Interlinking entities (URIs) using RDF



RDF triple:

```
<subject> <property> <object>
<the Mona Lisa> <was created by> <Leonardo da Vinci>
```

Source: https://www.w3.org/TR/rdf11-primer/

Linking Latin (LiLa)

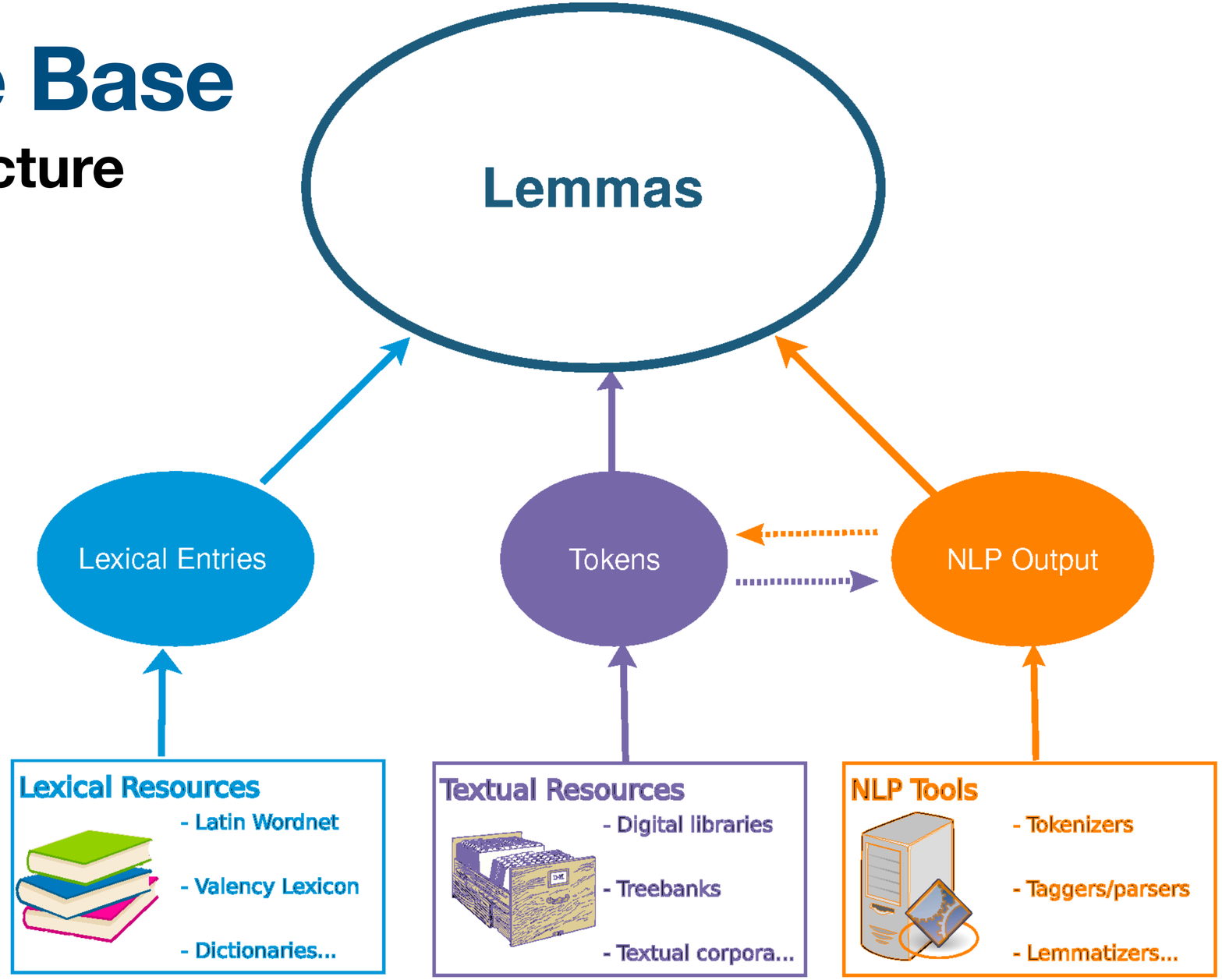
Linguistic Linked Open Data (LLOD) for Latin

ERC Consolidator Grant
2018–2023

10. 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.

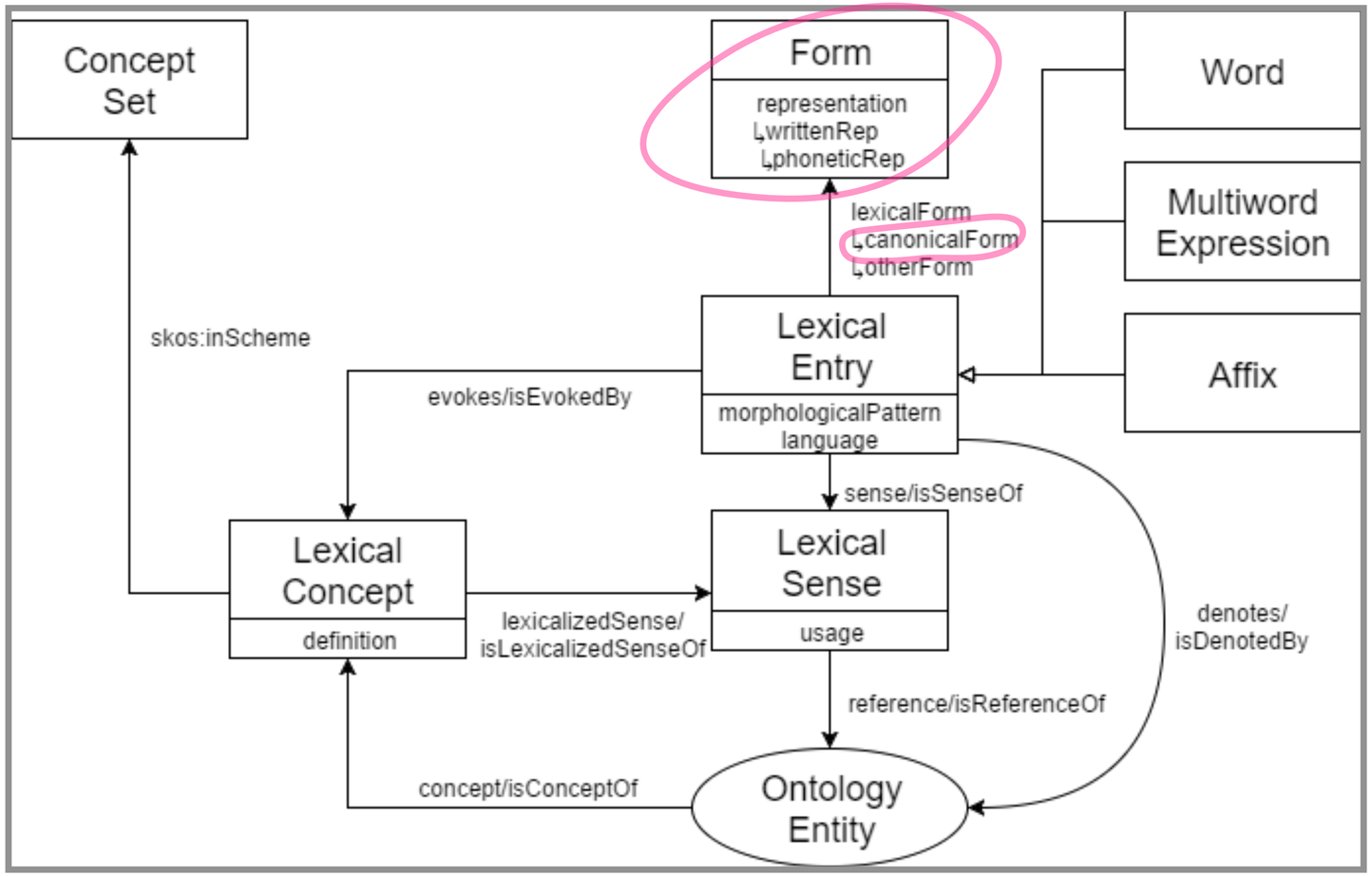
LiLa Knowledge Base

Lexically-based architecture and (meta)data sources



OntoLex-Lemon

A *de facto* W3C standard for publishing lexical data as LLOD



OntoLex-Lemon core model[11]

11. Cimiano, Philipp, 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/ [accessed 12 September 2024]

Lemmas

Using (and building upon) the OntoLex model

- LiLa/MOLOR Lemmas are modelled as a (subclass of) `ontolex:Form` (one URI for each lemma)
- OntoLex Forms have one or more **written representations** (as part of the same URI)
- Lexical entries in resources can be linked to lemmas with the `ontolex:canonicalForm` property
- Lemmas have at least a (single) POS

Lemma properties

An example of a LiLa lemma

- LiLa makes use of the property `lila:lemmaVariant` to distinguish between canonical forms that represent inflectional variants

hibernia

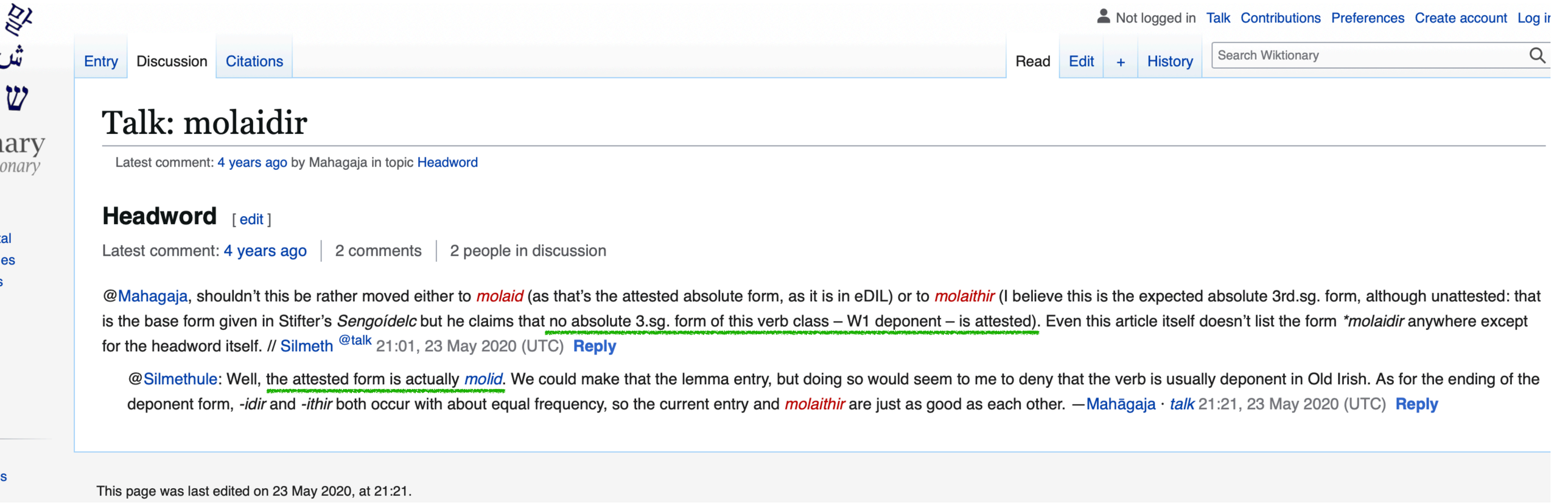
http://lila-erc.eu/data/id/lemma/9241

rdfs:**label**	hibernia
ontolex:**writtenRep**	hibernia @la iuberna @la iuerna @la
rdf:**type**	lila:Lemma ↳ Lemma
lila:**hasPOS**	lila:proper_noun ↳ proper noun
dcterms:**isPartOf**	<http://lila-erc.eu/data/id/lemma/LemmaBank> ↳ LiLa Lemma Collection
lila:**hasGender**	lila:feminine ↳ feminine
lila:**hasInflectionType**	lila:n1 ↳ first declension noun
lila:**lemmaVariant**	<http://lila-erc.eu/data/id/lemma/141861> ↳ ierne

Populating the MOLOR Lemma Bank: alignment procedures

The MOLOR lemma bank

Choosing a lemma in Old Irish …



Not logged in Talk Contributions Preferences Create account Log in

Entry Discussion Citations Read Edit + History Search Wiktionary

Talk: molaidir

Latest comment: 4 years ago by Mahagaja in topic Headword

Headword [edit]

Latest comment: 4 years ago | 2 comments | 2 people in discussion

@Mahagaja, shouldn't this be rather moved either to *molaid* (as that's the attested absolute form, as it is in eDIL) or to *molaithir* (I believe this is the expected absolute 3rd.sg. form, although unattested: that is the base form given in Stifter's *Sengoídelc* but he claims that no absolute 3.sg. form of this verb class – W1 deponent – is attested). Even this article itself doesn't list the form **molaidir* anywhere except for the headword itself. // Silmeth @talk 21:01, 23 May 2020 (UTC) **Reply**

@Silmethule: Well, the attested form is actually *molid*. We could make that the lemma entry, but doing so would seem to me to deny that the verb is usually deponent in Old Irish. As for the ending of the deponent form, *-idir* and *-ithir* both occur with about equal frequency, so the current entry and *molaithir* are just as good as each other. —Mahāgaja · *talk* 21:21, 23 May 2020 (UTC) **Reply**

This page was last edited on 23 May 2020, at 21:21.

source: https://en.wiktionary.org/wiki/Talk:molaidir

The MOLOR lemma bank

Provenance

- Current focus on nouns and verbs
- Corpus PalaeoHibernicum or CorPH (Stifter et al. 2021)[12]
  - 10k+ lemmas (also headwords from Greek, Latin, …)
  - Bulk of contemporary Old Irish material (but Würzburg absent!)
  - Deep morphological token-level annotation
- Goidelex (Anderson et al. 2024)[13]
  - 671 noun flexemes (< 574 noun lexemes) — entries based on Würzburg
  - Normalised orthography, used for automated phonemic transcriptions
  - Fine-grained and principled categorisation and annotation of morphosyntactic as well as phonological features (-> basis for a Paralex-compliant[14] inflected lexicon)

12. 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 1 October 2024]
13. Anderson, Cormac, Sacha Beniamine, & 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*, 1–10, Torino, Italia. ELRA and ICCL
14. Beniamine, Sacha, 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 [accessed 11 November 2024]

The MOLOR lemma bank

Provenance (cont'd)

- Würzburg lexicon (Kavanagh & Wodtko 2001)[15]
  - Headwords and attested forms for the Irish part of the Würzburg glosses + grammatical analysis
  - PDF files, experimentally parsed into csv format (not all POS)[16]
  - ~ 850 nouns (**574 in Goidelex**), ~ 316 verbs, …
- (electronic) Dictionary of the Irish Language or (e)DIL[17]
  - 43k+ headwords
  - 700–1700CE: Old Irish mixed with later periods/material — not clearly delineated — and many cross-references
  - retrodigitised into XML — yet often proves hard to navigate (and automatically parse!) due to editorial inconsistencies

15. 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
16. Prepared by Dr Aaron Griffith with help from the Utrecht Digital Humanities Lab (DHLab). The Würzburg glosses extraction scripts are at https://github.com/UUDigitalHumanitieslab/wurzburg-glosses-extraction [accessed 11 November 2024]
17. 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). <https://dil.ie/> [accessed 2 October 2024]

The MOLOR lemma bank

Nouns and mixed inflection classes

CorPH	Goidelex	eDIL (automatically parsed)[18]
fius, noun, u, n	**fius**, noun, u, n	**fis**, noun, u, o, n, m **fess**, x
	fius, noun, u, m	
	fius, noun, o, n	
breth, verbal_noun, ā, f	**breith**, verbal_noun, ā, f	**breth**, verbal_noun, ā, i, f brith, x breith, x
brith, verbal_noun, i, f	**brith**, verbal_noun, i, f	
fidbaid, noun, ā/i, f		**fidbaid** noun, ā, i, f
adaig, noun, ī, f	**adaig**, noun, ī1, f	**adaig**, noun, adverb, iā, f **aidche**, x **oidche**, x
pendaind, noun, ī, f	**pennait**, noun ī2*, f	**pennaind**, noun, [ ], f
pendait, noun, ī, f		**pennait**, noun, ā, f
aithis, noun, ā/ī, f	**aithis**, noun, ī2*, f	**aithis**, noun, [ ], f
eipistil, noun, ā/i/ī, f	**eipistil**, noun, ī2*, f	**epistil**, noun, [ ], f
talam, noun, n, m	**talam**, noun, n1, m	**talam**, noun, n, m
gein, noun, n, n	**gein**, noun, n2**, n	**gein**, verbal_noun, noun, [ ], f, n
persan, noun, ā/n, f	**persan**, noun, ā, f	**persa**, noun, n, f

*ī2, Mixed ī-/ā-stem, This flexeme is inflected as an ī-stem, with short acc.sg and dat.sg
**n2, n-stem noun, ā-stem genitive singular, This flexeme is inflected as an n-stem, except in the genitive singular, where it is inflected like an ā-stem

18. Parser available and adapted from https://github.com/ancatmara/crawlers/tree/main/DIL, created by Oksana Dereza [accessed 4 October 2024]

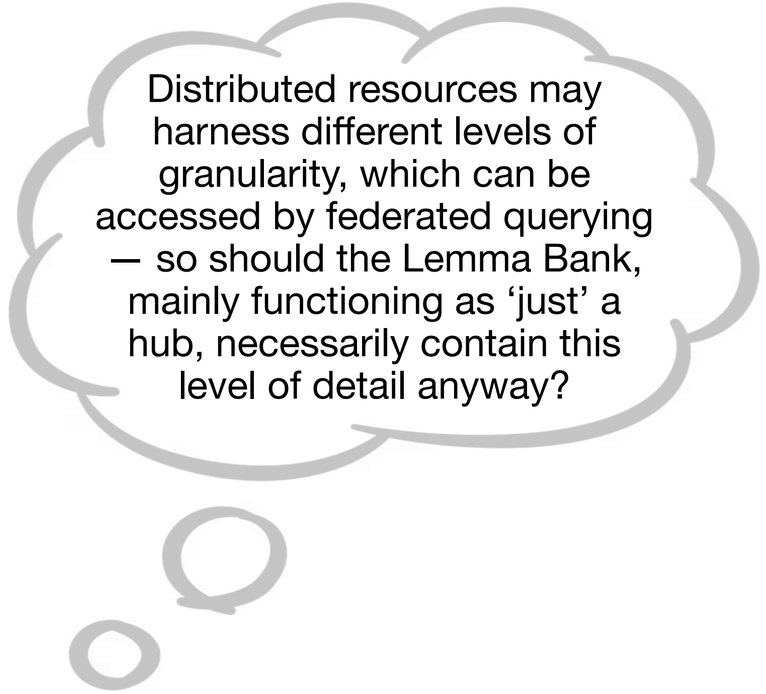
The MOLOR lemma bank

Alignment and harmonisation challenges with nouns

- Mixed/variable inflection classes (~5% of lemmas estimated based on Goidelex)
  - CorPH: 257 cases of *ī* and *n* inflection classes — some of which could be subcategorised — and 774 cases of *unk*/*?*/*NULL* for class/gender
- Granularity — different labels for inflection classes across resources (often microclasses)
  - e.g. *ī1*, *ī2*, *n1* and *n2*
- Inconsistent lemmatisation practices (1/more headwords per lexeme, cross-refs)
- Underlying: diachronic (synchronic?) variation as well as lack of attestation

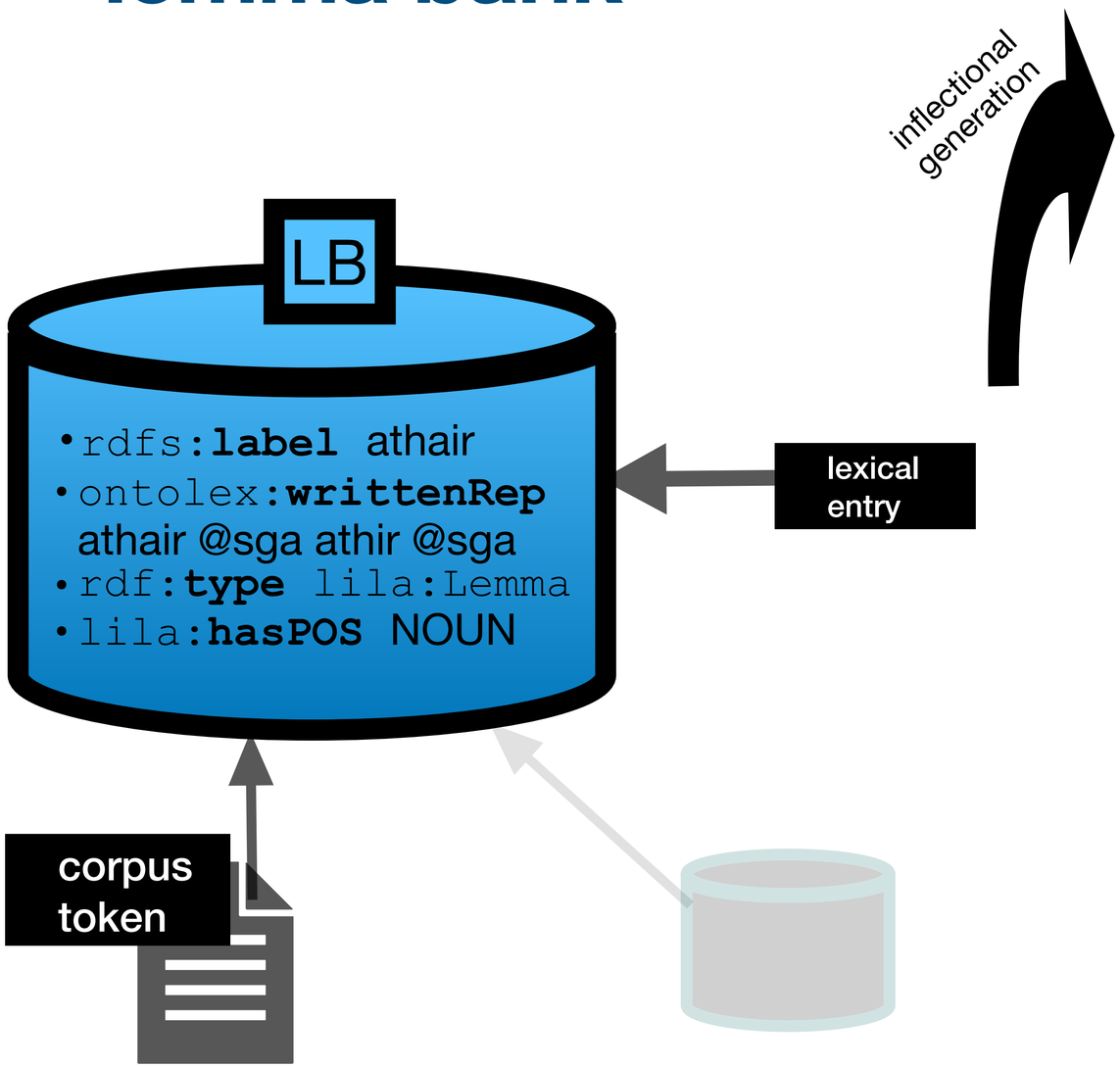
The MOLOR lemma bank

Solutions and statistics



- More minimal ontology for lemmas than in LiLa
- No inflectional information (for now)
- At least partly justified by often difficult-to-harmonise *granularity levels* and *editorial decisions* in the various available lexical resources for Early Irish
- **4750+** noun lemmas, **1100+** verb lemmas (mostly from CorPH)

The MOLOR *interlinked* lemma bank



lemma	class	tag	phono_lemma	morphono_lemma	phon_form	orth_trunc_p2g_form	orth_norm_p2g_form
athair	[Noun][Syncope][R][Masc]	[Nom][Sg]	ˈØʸaθʸərʲ	ˈØʸaθʸVrʲ	ˈØʸaθʸərʲ	athair;athir;atair;atir;ḟathair	athair
athair	[Noun][Syncope][R][Masc]	[Voc][Sg]	ˈØʸaθʸərʲ	ˈØʸaθʸVrʲ	ˈØʸaθʸərʲ	athair;athir;atair;atir;ḟathair	athair
athair	[Noun][Syncope][R][Masc]	[Acc][Sg]	ˈØʸaθʸərʲ	ˈØʸaθʸVrʲ	ˈØʸaθʸərʲ	athair;athir;atair;atir;ḟathair	athair
athair	[Noun][Syncope][R][Masc]	[Gen][Sg]	ˈØʸaθʸərʲ	ˈØʸaθʸVrʲ	ˈØʸaθʸarʸ	athar;atar;ḟathar;ḟatar	athar
athair	[Noun][Syncope][R][Masc]	[Dat][Sg]	ˈØʸaθʸərʲ	ˈØʸaθʸVrʲ	ˈØʸaθʸərʲ	athair;athir;atair;atir;ḟathair	athair
athair	[Noun][Syncope][R][Masc]	[Nom][Pl]	ˈØʸaθʸərʲ	ˈØʸaθʸVrʲ	ˈØʸaθʸərʲ	athair;athir;atair;atir;ḟathair	athair
athair	[Noun][Syncope][R][Masc]	[Voc][Pl]	ˈØʸaθʸərʲ	ˈØʸaθʸVrʲ	ˈØʸaθʸrʸaØʸ	athra;atra;ḟathra;ḟatra	athra
athair	[Noun][Syncope][R][Masc]	[Acc][Pl]	ˈØʸaθʸərʲ	ˈØʸaθʸVrʲ	ˈØʸaθʸrʸaØʸ	athra;atra;ḟathra;ḟatra	athra
athair	[Noun][Syncope][R][Masc]	[Gen][Pl]	ˈØʸaθʸərʲ	ˈØʸaθʸVrʲ	ˈØʸaθʸrʸaØʲ	athrae;athre;athra;atrae;atre	athrae
athair	[Noun][Syncope][R][Masc]	[Dat][Pl]	ˈØʸaθʸərʲ	ˈØʸaθʸVrʲ	ˈØʸaθʸrʸəβʲ	athraib;athrib;atraib;atrib;ḟathraib	athraib
athair	[Noun][Syncope-j][R][Masc]	[Nom][Sg]	ˈØʸaθʸərʲ	ˈØʸaθʸVʲrʲ	ˈØʸaθʸərʲ	athair;athir;atair;atir;ḟathair	athair
athair	[Noun][Syncope-j][R][Masc]	[Voc][Sg]	ˈØʸaθʸərʲ	ˈØʸaθʸVʲrʲ	ˈØʸaθʸərʲ	athair;athir;atair;atir;ḟathair	athair
athair	[Noun][Syncope-j][R][Masc]	[Acc][Sg]	ˈØʸaθʸərʲ	ˈØʸaθʸVʲrʲ	ˈØʸaθʸərʲ	athair;athir;atair;atir;ḟathair	athair
athair	[Noun][Syncope-j][R][Masc]	[Gen][Sg]	ˈØʸaθʸərʲ	ˈØʸaθʸVʲrʲ	ˈØʸaθʸarʸ	athar;atar;ḟathar;ḟatar	athar
athair	[Noun][Syncope-j][R][Masc]	[Dat][Sg]	ˈØʸaθʸərʲ	ˈØʸaθʸVʲrʲ	ˈØʸaθʸərʲ	athair;athir;atair;atir;ḟathair	athair
athair	[Noun][Syncope-j][R][Masc]	[Nom][Pl]	ˈØʸaθʸərʲ	ˈØʸaθʸVʲrʲ	ˈØʸaθʸərʲ	athair;athir;atair;atir;ḟathair	athair
athair	[Noun][Syncope-j][R][Masc]	[Voc][Pl]	ˈØʸaθʸərʲ	ˈØʸaθʸVʲrʲ	ˈØʸaθʲrʲaØʸ	aithrea;aithre;aithra;aitrea;aitre	aithrea
athair	[Noun][Syncope-j][R][Masc]	[Acc][Pl]	ˈØʸaθʸərʲ	ˈØʸaθʸVʲrʲ	ˈØʸaθʲrʲaØʸ	aithrea;aithre;aithra;aitrea;aitre	aithrea
athair	[Noun][Syncope-j][R][Masc]	[Gen][Pl]	ˈØʸaθʸərʲ	ˈØʸaθʸVʲrʲ	ˈØʸaθʲrʲaØʲ	aithre;aitre;athre;atre;oithre	aithre
athair	[Noun][Syncope-j][R][Masc]	[Dat][Pl]	ˈØʸaθʸərʲ	ˈØʸaθʸVʲrʲ	ˈØʸaθʲrʲəβʲ	aithrib;aitrib;athrib;atrib;oithrib	aithrib

Conclusions and future work

Conclusions

- A Lemma Bank as a hub in a Linguistic Linked Open Data architecture for Old Irish
  - Benefits: controlled vocabularies, semantic web standards, FAIR, etc.
  - Interlinking leads to discovery (-> didactic opportunities)
- Design issues:
  - Variable morphological inflection due to e.g. lack of attestation/uncertainty and inflectional variation
  - Reflected in the lexicographical resources
- Solution: adopting a more minimal ontology for lemmas than in LiLa
  - But: possibility to expand at a later stage

Future work

- Short-term:
  - Lemma Bank: more POS categories (at least adjective) and RDF conversion
  - Interlinking of one or two resources (time permitting)
  - Lemma augmentation through alignment/extraction with/from eDIL (time permitting)
- Long-term:
  - More (mature) NLP tools for historical Irish (Goidelic) languages (e.g. lemmatiser, morphological analyser, …) for optimal exploitation of Lemma Bank and LLOD for Early Irish in general
- Life-goal (?)
  - Diachronic, pan-Gaelic NLP tool development and resource interlinking (“GLTK”[19]? 😀)

19. Analogous to CLTK: The Classical Language Toolkit (http://cltk.org/)

Thank you

Questions?

MOLOR: Morphologically Linked Old Irish Resource
Università Cattolica del Sacro Cuore
CIRCSE Research Centre

theodorus.fransen@unicatt.it

https://tinyurl.com/molor-project

@ThFransen

Largo Gemelli 1, 20123, Milan, Italy

UNIVERSITÀ CATTOLICA del Sacro Cuore

CIRCSE
Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione

Funded by the European Union

This project has received funding from the European Union's Horizon Europe scientific research initiative under the Marie Skłodowska-Curie Actions (MSCA) - grant agreement no. 101106220