# BUILDING A LEMMA BANK FOR OLD IRISH VERBS

Federico Simone Samperi (Università di Pavia)

6th Pavia International Summer School for Indo-European Linguistics , 2-7 September 2024

## INTRODUCTION

This poster is aimed at describing the on-going process towards building a Lemma Bank for Old Irish verbs. My work is part of Dr Theodorus Fransen's (UCSC) larger Marie Curie project MOLOR_db, a Lemma Bank for Old Irish (600-900CE). The Lemma Bank is being built within the LLOD (Linguistic Linked Open Data) paradigm and it is ispired by a similar resource for Latin, designed as part of the LiLa project (2018-2023) (Fransen 2023:37).

## WHY A LEMMA BANK FOR OLD IRISH VERBS?

Unlike other ancient languages (for ex. Latin) , Old Irish has a fairly recent history (600-900CE). When dealing with Old Irish one must consider:

- Lack of orthographic standardisation
- Scarsity in comprehensive resources
- Different lemmatisation standards
- Very large corpus of texts but not entirely translated or even annotated

The aim is to **create a comprehensive lexical resource using the LLOD framework** (LiLa project principles).

## DATA AND METHODOLOGY

Lemmas were systematically collected and stored in an Excel worksheet from the following lexical resources:

1. **Kavanagh** (Kavanagh, Séamus, and Dagmar S. Wodtko 2001)
2. **Corpus Paleo Hibernicum** (CorPH, Stifter 2021)
3. **Electronic Dictionary of the Irish Language online** (eDIL)

### STEP 1: DATA COLLECTION

(A) Verbs from CorPH were extracted automatically beforehand and placed in alphabetical order on an Excel sheet.

(B) All corrisponding lemmas from the other two resources were manually searched, checked and added one by one to the file.

### STEP 2: DATA SYSTEMATIZATION

Every form encountered was either labelled as:

- **Lemma**
- **Lemma variant:** canonical forms with morphological/inflectional differences.
- **Spelling variant:** orthographical differences ("written representation", Ontolex model 2016).

**Lemma variants were modelled on a different row but in the same column ; spelling variants were modelled on the same row but in a different column (namely wr2, wr3, wr4 etc.).**

### STEP 3: PROPERTY ASSIGNMENT

(A) To each form a code (MOLOR_id) was assigned.

An integer number for lemmas and spelling variants; a decimal number for lemma variants.

(B) Each CorPH form was associated with a CorPH_id, modelled in a separate column and every eDIL form was associated to a link.

### STEP 4: OTHER INFORMATION

(A) Translation was given (usually either from CorPH or eDIL)

(B) POS tag VERB was assigned

(C) Inflectional class was assigned using **McCone's standard (1987)**

(D) Etymology was retrieved from various sourced and added

## RESULTS

Distribution of the forms in the dataset: CorPH **42%** ; Kavanagh **13%** ; eDIL **45%**.

- Total forms collected: **2597**
  Total lemmas + lemma variants: **1284**
  Total number of lemmas: **1125**
  Total number of lemma variants: **159**
  - with an INFL_CLASS assigned: **1261**
  - w/out an INFL_CLASS assigned: **23**
- Three new lemmas only found in Kavanagh (contained in the Wb Glosses): **caín-airlethar** "to take good heed"; **ceta-creti** "to believe for the first time"; **cetu-pridcha** "to first pray".
- CorPH and Kavanagh use the 3sg. to lemmatise, whereas 7 forms for eDIL are lemmatised either using the 1sg. (6 forms) or the 1pl. (1 form)
- New Lat. loans detected: **áctegim** (<acetum); **iúdigid** (<iudaizare)

## FUTURE AIMS

- Add all the other lemmas which are not listed as headword(s) in eDIL
- Detect diachronic variation within the dataset, separating Old vs Middle Irish forms
- Interlinking (LiLa for Latin derived verbs; Pa.Ve.Da)
- Correct/Add missing information (i.e INFL_CLASS, etymology)

### ACKNOWLEDGEMENTS AND CONTACTS

Special thanks to Dr. Theodorus Fransen, my supervisor at CIRCSE (UCSC) and M.A thesis co-supervisor, for sharing his project with me and for allowing me to work with him. I would also like to thank Prof. Elisa Roma for helping me with my master's thesis. Feel free to contact us at:
federicosimone.samperi01@universitadipavia OR theodorus.fransen@unicatt.it

**Scan the QR code to see the bibliography**