Guideline for the structure of the CorPH database

latest version: 30/11/2020

This will be updated!!

1. Introduction

CorPH stands for *Corpus Palaeohibernicum* 'a corpus of Old Irish'. It was constructed by the ChronHib project between 2015 and 2020. CorPH aims to provide a solid basis for study of the Old Irish language in the form of an electronic, linguistically annotated corpus of Old Irish texts. It incorporates earlier lexical databases (see Milan glosses, St. Gall glosses, The poems of Blathmac) as well as new data digitised and annotated by the ChronHib team. For an overview of the criteria by which texts are chosen to be incorporated into CorPH, please see here.

CorPH is stored in two formats: a relational database accessible via link, and textual files that contain individual annotated texts (link). This document explains the structure of the former.

2. Basic structure

The CorPH relational database consists of the following tables:

- Morphology
- Lemmata
- Sentences
- Texts

Each table consists of columns (attributes) and rows (values) with a unique ID identifying each row. Below the columns of each table are explained.

3. The 'Morphology' table

This table is the core of the database, and stores most of the linguistic information. It consists of the following columns:

ID_unique_number

> This is a unique key for every token (row) in the table. Please note that the numerical order of the ID_unique_number does not necessarily follow the sequence of words in the texts.

Textual_Unit_ID

> This denotes the textual unit in which the 'morph' is found. It is linked to the column with the same name in the 'Sentences' table.

TextID

This denotes the text in which the ‘morph’ is found. It is linked to the column with the same name in the ‘Texts’ table.

Stressed_Unit

For the definition of the ‘stressed unit’ see the Guidelines for stressed units, morphs and expected morphs.

Morph

For the definition of the ‘morph’ see the Guidelines for stressed units, morphs and expected morphs.

Expected_Morph

For the definition of the ‘expected morph’ see the Guidelines for stressed units, morphs and expected morphs.

Lemma

This column stores the lemma, or dictionary headword, to which the morph belongs. It is linked to the column with the same name in the ‘Lemmata’ table.

Secondary_Meaning

If the morph has a specific or derived meaning in the context that is not generally used or is not captured by the ‘Meaning’ of the lemma in the ‘Lemmata’ table, such as ‘substantivised’ (for adjectives), ‘the Son, i.e. Jesus (for the lemma “macc 1”)’ or ‘used as a personal name’ (for adjectives and nouns), etc., the special meaning is entered here.

Analysis

This column stores the morphological annotation tags, see the Guideline for morphological tagging.

Relativity, Transitivity, Dependency, Deponent, Contraction, Augment, Hiatus, Mutation, Causing_Mutation, Hybrid_form, Problematic_Form, Onomastic_Complex, Onomastic_Usage

For the values of these columns see the Guideline for morphological tagging. They are not currently used across the entire corpus.

Comment

This column is reserved for any other comments that do not fit into any of the abovementioned columns.

Syntactic_ID, Phrase_structure_tree, Syntactic_Unit, Phrase_type, Phrase, Syntactic_Unit_Translation

For the values of these columns see the Guideline for syntactic tagging. They are not currently used across the entire corpus.

ID_of_Change

This column stores the variation tags, see the Guideline for variation tagging for more information.

4. Lemmata

This table stores information about the lemmata (headwords), and can be used as a lexicon. It consists of the following columns:

ID_unique_number

This is a unique key for every token (row) in the table.

Lemma

This column stores the lemma, or dictionary headword. It is linked to the column with the same name in the 'Morphology' table. The lemma form is generally spelled according to the standardised Old Irish phonology and orthography, i.e. with glide letters <a> <i> <e>, <h> only after <t> <c>, <óe> and <oí> for original /oi̯/, etc. For special lemma names for pronouns, particles and Latin words that have Irish homonyms, see the Guideline for Lemmata tagging.

Meaning

This column provides the English equivalent to the most common or general meaning of the lemma. For verbs, the infinitive 'to X' is used. For proper names, the literal meaning, where this is known, is put between quotation marks, e.g. 'small-fire' for *Áedán*.

Part_Of_Speech, Classification, Gender, Etymology, Language

For the values and tags of these columns, see Guideline for Lemmata tagging.

Comments

This column is reserved for any other comments that do not fit into the abovementioned columns.

DIL_Headword

This column provides a URL link to the most relevant eDIL entry.

5. Sentences

This table has now been renamed 'Textual_unit' in the Upfront website, as it contains different types of textual units that are divided according to conventions found in different editions: individual glosses in collection of glosses, sentences in prose texts, stanzas in verses, columns in lists of names, etc.

ID_unique_number

This is a unique key for every token (row) in the table. Please note that this does not necessarily follow the sequence of the textual units in a text.

TextID

This denotes the text in which the ‘morph’ is found. It is linked to the column with the same name in the ‘Texts’ table.

Textual_Unit_ID

This is a unique key for every textual unit in the sequence of their occurrence in a text. The format is in the form of SXXXX-XXXX (X being a number), where SXXXX is the same as the TextID of the text in which it is found, e.g. S0011-24 is the 24th textual unit it text S0011. This column is linked to the column with the same name in the ‘Sentences’ table.

Locus1, Locus2, Locus3

In these columns one finds the conventional sigla or numbers for the textual units in previous editions, e.g. Ml. 002a05, A 20ra, Thes.ii.44.P.1,1.6, etc. There are three columns because more than one numbering system may exist in different editions.

Textual_Unit

This column contains the text as found in the source edition of the textual unit in question.

Subunit

This column is currently not used.

Translation

This column provides an English translation of the textual unit.

Latin_Text

This column provides the Latin text to which a gloss refers. It is only used for texts that are Old Irish glosses on Latin texts.

Latin_Translation

This column provides an English translation of the Latin_Text. It is only used for texts that are Old Irish glosses on Latin texts.

Textual_Notes

This column provides any textual notes or comments by previous editors on the textual unit.

Translation_Notes

This column provides any notes or comments on the translations of the textual unit or the Latin text.

Variant_Readings

If there are variant readings of the textual unit from other copies, they are found in this column.

6. Texts

This table lists the texts included in CorPH and metadata about their dates, locations, etc.

ID_unique_number

This is a unique key for every token (row) in the table.

Text_ID

This is also a unique key for every text, in the format of SXXXX (where X is a number), e.g. S0022. It is linked to the column with the same name in the ‘Sentences’ and ‘Morphology’ tables.

Title_of_text

This column contains the names of the texts, but is not obsolete and not displayed on the Upfront.

Revised_title

This column contains the revised, more unified names of the texts, and is now displayed on the Upfront.

Mss

This column lists the manuscripts in which the text is found.

Digital_mss

This column gives the URL link to the digital manuscript images, if available.

Date

This column contains the known or estimated date of the manuscript and the date of the text (if they differ). The dates are usually expressed in range in the format of ‘825-850’.

Edition

This column lists bibliographic information about previous editions of the text.

Dating_criteria

This column includes a brief overview of the information relevant to the dating of the manuscript and the text.

Thes

Since a number of texts have been edited in *Thesaurus Palaeohibernicus*, the page and line number in the *Thesaurus* are quoted here.

Worker

This column records the person who is mainly responsible for processing the text data: EL = Elliott Lash, BB = Bernhard Bauer, FQ = Fangzhe Qiu, SB = Siobhán Barrett, RB = Romanas Bulatovas, LN = Lars Nooij, DS = David Stifter, AG = Aaron Griffith.

Created_date

This is the date when the text data were uploaded to CorPH.

Checked_against_the_mss

This indicates whether the digitised text has been checked against the manuscript or manuscript images.

Reason_of_MS_choice_and_editorial_policy

This column is now obsolete, and a separate document is available which explains the reason of our choice of manuscripts as source for data and our editorial policy.