**Lemmata tagset for the CorPH database**

Fangzhe Qiu, Elliott Lash, Bernhard Bauer, David Stifter

latest version: 29/11/2019

old

**1. Introduction**

This document details guidelines for tagging the various attributes in the Lemmata table in the CorPH database. For the structure of the database please refer to the Guideline. These attributes include ‘Lemma’, ‘Meaning’, ‘Part\_Of\_Speech’, ‘Classification’, ‘Gender’, ‘Etymology’, ‘Language’, ‘Comments’ and ‘DIL\_Headword’. They are all attributes describing aspects of the lemma, which is the key value in this table and therefore is unique. This document only deals with the following attributes: ‘Part\_Of\_Speech’, ‘Classification’, ‘Gender’ and ‘Language’.

For each attribute, a limited set of tags are used. These tags do not contain blank space but may contain underlines (\_).

**2. POS tags for CorPH**

For an introduction to the word categories of Old Irish, see Stifter, David, ‘Early Irish’, in Müller, Nicole and Ball, Martin J. (ed.) *The Celtic languages*, 2nd edition, (Routledge 2009), pp. 55–116. A more fine-grained tagset is used here to increase the linguistic accuracy, and the full description of the part-of-speech is used instead of abbreviations to facilitate query. The tag names are built ‘hierarchically’, so that additional words in the name of a tag specify a type of the superordinate category. If a lemma is tagged with a POS from a sub-category, it is not tagged with a POS from a higher category, e.g. *Fergus* is tagged with ‘proper\_noun’ and not with ‘noun’.

A problem arising from regarding the POS as an attribute of the lemma rather than of the word in the context is that, according to the syntactic function the word has in the sentence, the POS may change. For instance, an adjective may be substantivized to serve as the head of a noun phrase. Fortunately for Old Irish, the POS of a word is relatively stable and the change in POS is mostly achieved by derivation. Therefore, the POS of a lemma is usually easy to determine through the most common function it serves in sentences. When the syntactic function of a word in the sentence requires a different POS than that which has been tagged to the lemma, such information is recorded in the ‘Morphology’ table in the column of **Secondary\_Meaning**, as well as in the **syntax** field. Substantivized adjectives are tagged as the head of NP (noun phrase) not AdjP, for example.

The fine-grained POS tags also help to capture the variety of POS some lemmata frequently show in different syntactic contexts, e.g. *óen* ‘one, single, one (thing or person)’ is given the POS tag **adjective\_numeral\_noun**.

The POS tags used in CorPH are as follow:

* **adjective**

This tag is for all adjectives apart from the special subcategories below.

* **adjective\_numeral**

This tag is for *dá* ‘two’, *trí* ‘three’ and *cethair* ‘four’ which inflect.

* **adjective\_numeral\_noun**

This tag is for *óen* ‘one’ and the ordinal numerals which inflect like o/ā adjectives but frequently mean ‘a third, the third day, three-day period’ etc.

* **adjective\_quantifier**

This tag is for unstressed *cach* ‘every’, *nach* 1‘any’, etc. and *uile* ‘all,

each’.

* **adjective\_pronominal**

This tag is for *aile*, *alaile*, *ala* ‘other, another’.

* **adverb**
* **complementiser**

This tag is for *co* 3, *co* 4 ‘so that, until’, *ara* ‘so that’, *día* 1 ‘if’.

* **conjunction**
* **definite\_article**
* **interjection**
* **noun**

This is the tag for all nouns apart from the special sub-categories below.

* **noun\_numeral**

This tag is for the decades, *cét* 1 ‘a hundred’ and *míle* ‘thousand’.

* **proper\_noun**

This tag is for the names of unique/specific persons, places, languages and spiritual entities.

* **verbal\_noun**

This tag is for all nouns that are derived from a verbal root. Abstract nouns form which denominative verbs are derived (GOI §722), e.g. *rím* ‘counting, number’ > *rímid* ‘to count’, are not regarded as verbal nouns.

* **numeral**

This tag is for the ‘true’ numerals 5 to 9 that do not inflect.

* **\*particle**

The tag ‘particle’ does not exist on its own, but various particles are tagged in the following ways:

* **particle\_augment**

This tag is for *com*·, *con*· 6, *ro*· 2, and *ad*· 2.

* **particle\_anaphoric**

This tag is for clitics *-side*, -*són* and *-ón* ‘that aforementioned’.

* **particle\_comparative**

This tag is for *oldaäs*, *indaäs,* *adaäs* ‘than’. Although these are regarded as phrases in CorPH and are further divided into the prepositions, the neuter article, the nasalising relative particle and the substantive verb, adding an extra morph and lemma for the whole phrase facilitates the study of this structure.

* **particle\_demonstrative\_proximate**

This tag is for clitics *-se* and *-so­* ‘this’.

* **particle\_demonstrative\_distal**

This tag is for clitics *-sin* ‘that’and *-ucut* ‘that (over there)’.

* **particle\_discourse**

This tag is for discourse particles *danó*, *didiu*, *immurgu* etc.

* **particle\_focus**

This tag is for *acht* 1 ‘but’, *céin* ‘…too’, etc..

* **particle\_interrogative**

This tag is for *in* 2.

* **particle\_negative\_main**

This tag is for the negative particles *ní* 3, *ná* 1, *nach* 2, *nicon* that heads a main clause.

* **particle\_negative\_subordinate**

This tag is for the negative particles *ná* 2, *nád* 1, *nach* 3, *nacon* and *nádcon* that heads a subordinate clause.

* **particle\_numerative**

This tag is for the particle *a* 3.

* **particle\_prefix**

This tag is used for prefixes of various meaning, e.g. privative, intensive, first-, mid-, etc., but is not used for adjectives (or adjectival numerals) that can exist independently, e.g. *óen*, *sen, óg* etc.

* **particle\_preverb**

This tag is for preverbs that modify the lexical or aspectual meaning of the

verb but do not include augments.

* **particle\_pronominal**

This tag is for the *nota augens*.

* **particle\_relative**

This tag is for the infixed leniting and nasalising relative particles.

* **particle\_vocative**

This tag is for *a* 1 ‘oh’.

* **preposition**
* **\*pronoun**

The tag ‘pronoun’ does not exist on its own, but various pronouns or pronoun- like items are tagged in the following ways:

* **pronoun\_anaphoric**

This tag is for the stressed *suide* 1 and *sodain*.

* **pronoun\_demonstrative\_distal**

This tag is for the stressed *sin*.

* **pronoun\_demonstrative\_proximate**

This tag is for the stressed *só*, *sé* 2.

* **pronoun\_indefinite**

This tag is for *sechi* and *cacha* ‘whoever, whatever’.

* **pronoun\_independent**

This tag is for the stressed independent personal pronouns, including the genitive forms *muí*, *taí, aí, náthar*, etc.

* **pronoun\_infixed**
* **pronoun\_interrogative**

This tag is for pronouns *cía* 1 ‘who’, *cid* 1 ‘what’, *cote* ‘what is’ etc.

* **pronoun\_possessive**

This tag is for the possessive pronouns.

* **pronoun\_propword**

This tag is for *í* ‘that one/thing’

* **pronoun\_quantifier**

This tag is for *cách* ‘each, every’, *ní* 1 ‘something’, *cechtar* ‘each of two,

either’, etc.

* **pronoun\_emphatic**

This tag is for *fadéin*, *féin*, ‘that one, oneself’, etc. (GOI §485).

* **pronoun\_relative**

This tag is for the relative pronoun *–(s)a* ‘that which’ used in prepositional

relative clauses.

* **pronoun\_suffixed**
* **verb**
* **verbal\_of\_necessity**

This tag is for the gerundive verbal form described in GOI §717-9.

* **verbal\_participle**

This tag is for the past passive participles formed with the suffix -*tii̯o/ā* to a verbal root (GOI §714-6), even when the underlying verb is not attested.

**3. ‘Classification’ tags for CorPH**

The ‘Classification’ column stores information of several types:

3.1 Adjectives and nouns (including their sub-categories and the verbal\_participle) are tagged with their stem types:

**o ā o/ā i̯o i̯ā i̯o/i̯ā i ī u**

**k g t d nt n r s**

**irreg. indecl. unk.**

The tag **irreg.** is for nominal whose paradigm does not follow the regular patterns of the other nominal stems in Old Irish, e.g. *bó* ‘cow’.

The tag **indecl.** is for nominal which is indeclinable, e.g. *Ísu* ‘Jesus’.

The tag **unk.** is for cases where the stem type is unknown.

Multiple possibilities of values are separated by ‘**/**’, e.g. *eclais* ‘**ā/i/ī**’.

A doubtful value is followed by ‘**?**’.

3.2 Verbs are tagged with their inflectional classes (using the system in McCone, Kim, *The Early Irish Verb*, 2nd ed. (Maynooth, 1997)):

**W1 W2a W2b**

**S1a S1b S1c S2 S3a S3b**

**H1 H2 H3**

**copula defective**

3.3 Prepositions are tagged with the cases they govern:

**acc. dat. gen. acc./dat.**

**4. ‘Gender’ tags for CorPH**

Three basic tags are used:

**masc. neut. fem.**

In the case of uncertain or multiple genders, all plausible possibilities are listed in the order given above, separated by with slashes ‘**/**’, e.g. **masc./neut./fem.**.

**5. ‘Language’ tags for CorPH**

When the lemma is not Irish, it is tagged with the following codes in the ‘language’ column:

**Latin Greek Hebrew**

**British Pictish**

**Old\_English Old\_Norse Old\_High\_German Frankish**

These lemmata belong mostly to words in texts that show code-switching, but they also include contemporary foreign personal names that are gaelicised in spelling (e.g. *Amlaíb* < ON *Áleifr*). They do not, however, include loanwords from other languages into Irish, biblical names, and foreign names that were borrowed into Irish at an earlier stage and became established as a native name, e.g. *Pátraic, Brénainn, Afraic*, etc..