Automatic Morphological Parsing of Old Irish Verbs Using Finite-State Transducers

Theodorus Fransen
Data Science Institute, National University of Ireland Galway
E-mail: theodorus.fransen@nuigalway.ie

Abstract

The topic of this paper constitutes the main part of a recently finished Ph.D. project carried out by the author which investigates how computational methods can be employed to map cognate verb forms in Early Irish (ca. 7th–12th centuries A.D.) and Modern Irish (ca. 1200 onwards). This paper discusses the development of a finite-state morphological transducer using *foma* (Hulden, 2009) for the Old Irish language (ca. 7th–9th centuries A.D.), focusing on verbs. Two main challenges are discussed. First, different practices of word segmentation have significant repercussions for the encoding of dependencies both on and beyond the word level. A second challenge is complex verb stem formation and considerable stem allomorphy. This has been tackled by operating with "monolithic stem" entries for each verb lemma, i.e., synchronic, invariable hard-coded stems, representing a semi-surface-level base form.

Keywords: Old Irish verbs; computational morphology; finite-state transducers; stem allomorphy; word segmentation

1 Introduction

The creation of a morphological parser for Old Irish verbs was part of the author's Ph.D. project, the aim of which was to design a computational architecture to better facilitate systematic linguistic study of the historical development of the Irish verb. A major objective was (and is) to integrate available digital tools into this architecture, and create links between them, in order to create mappings between etymological cognates (Fransen, 2019; Fransen, 2020).

Old Irish, dated to ca. 7th–9th centuries A.D. (Russell, 2005), is "the earliest period of Irish—or of any Celtic language—for which the extant record is sufficiently full and varied to permit a full synchronic description" (Stifter, 2009, p.59). The standard work *A Grammar of Old Irish* (*GOI* = Thurneysen, 1946) is almost entirely based on the language of interlinear and marginal glosses found in Latin texts; these glosses are contemporary to the Latin manuscripts and can be safely assigned, unlike other medieval texts, to the Old Irish period. The language of the 8th and 9th centuries A.D., which we know mainly on the basis of the Old Irish glosses, is sometimes referred to as Classical Old Irish (see, e.g., Russell, 2005, p.407).

Old Irish can be clearly differentiated from Middle Irish (ca. 10th–12th centuries A.D.), which shows "a far-reaching overhaul of the verbal system" (McCone, 1997, p.165). This is a major challenge in the author's project, but not of major relevance for the purposes of this paper. Early Irish is often used

to denote Old and Middle Irish, i.e., the Irish language of the early medieval period. From roughly 1200 A.D. onwards, one speaks of Modern Irish.

There is a lack of digital resources for historical Irish to create links, in a systematic manner, between medieval and modern cognates. Recent advances in Natural Language Processing have mostly been made in the context of Modern Irish. Uí Dhonnchadha et al. (2014) describe how a POS-tagger (with a morphological finite-state transducer backbone) for the modern contemporary language has been augmented with a standardiser[1] to recognise and successfully process non-standard, pre-standard and increasingly earlier modern forms. However, for the medieval period no automatic morphological analysis tools currently exist. To bridge the hiatus in computational support, the author's focus is on automatic morphological analysis for Early Irish, more specifically Old Irish, as this period a) is relatively well-resourced, b) shows a relatively stable grammar and orthography,[2] at least when compared to Middle Irish, and c) "furnishes a yardstick with which to assess the abundant literary production of the medieval period" (Stifter, 2009, p.59).

The rest of this paper is structured as follows. A survey of important literature and already existing computational resources is presented in section 2. Section 3 provides a short description of the Old Irish verb. The computational paradigm of finite-state morphology is the subject of section 4. Section 5 details some aspects of the implementation, with a focus on word segmentation issues and the encoding of stems. A selection of the results is presented in section 6. Sections 7 and 8 provide a discussion and conclusion, respectively.

2 Computational Resources for Early Irish

Using NLP to deal with language variation in historical texts is far from straightforward:

> There is no underlying computational model that describes how synchronic and diachronic variants relate to each other and—possibly—to some shared meaning or some

[1] An Caighdeánaitheoir "The Standandardiser", developed by Kevin Scannell. See https://github.com/kscanne/caighdean and https://cadhan.com/ [accessed 19 December 2020].

[2] For a discussion on the two orthographic systems in the earliest Old Irish glosses, their diachronic dimension and their features, see Ó Cróinín (2001). Old Irish texts are commonly believed to show little or no trace of synchronic variation (Stifter, 2009, p.60). McCone (1985) cites various variant phenomena in the glosses, which he attributes to "lapses" from an educated register (i.e., stylistic variation). Recent work on the sociolinguistics of Old Irish and the question of dialect (i.e., geographic variation) in this period is Ó Muircheartaigh (2015). At the 2018–2019 O'Donnell Lecture at the University of Oxford (10 May 2019), Professor David Stifter expressed views that contradict some dominant scholarly views on the existence of a literary standard in the Old Irish period. His ERC-funded Chronologicon Hibernicum project has found that there is much more linguistic variation within Old Irish than is commonly assumed; some of this synchronic variation may be diatopic or diastratic. Moreover, according to Stifter, traditional statements suggesting the existence of a literary standard show a partial neglect of the sociolinguistic implications of a standard text language spread over a vast area.

kind of prototype that represents the relatedness of the variants (Piotrowski, 2012, p.9).

This problem, which applies to most, if not all, historical languages, is compounded by the fact that for historical Irish we have many disparate projects each focusing on a restricted time frame. This is not the place to discuss each individual resource in detail, and the reader is referred to the author's Ph.D. thesis (Fransen, 2019), which includes a survey (in Appendix A) of available digital linguistic resources (as well as dormant projects) for historical Irish. This survey has shown that there is a "lexicographical gap" between Old and Modern Irish, which lies at the heart of the author's thesis and future research aspirations. The remainder of this section discusses resources for Old and Middle Irish (i.e., Early Irish), which are the most relevant ones for the purposes of the present paper, which specifically deals with automatic morphological parsing of Old Irish verbs.

The XML-encoded *electronic Dictionary of the Irish language* or eDIL (Toner et al., 2019), covering the period ca. 700 A.D.–ca.1700, is the most authoritative and indeed standard dictionary for historical Irish. However, varying editorial practices across fascicles constituting the original printed DIL resulted in inconsistencies that found their way into the retro-digitised edition, for example with the spelling of headwords. Moreover, the dictionary is not comprehensive in terms of the inflected forms of headwords provided. It should be added, however, that the original objective of the eDIL project was not to revise the original hard-copy dictionary, but to open up the wealth of information contained in it and to make it accessible to a variety of users (Fomin and Toner, 2006).

Dereza (2018), who discusses lemmatisation approaches for ancient and morphologically complex languages, reports on lemmatisation strategies for Early Irish. She decided to avoid rule-based approaches involving stem and affixes and statistical machine learning methods due to the purported morphophonological complexity, non-transparent orthographic features and scarcity of data of the language stages in question. She resorted to an already existing dictionary, i.e., eDIL, from which she extracted *form:lemma* mappings while comparing two methods: 1) a lemma predictor based on the Damerau-Levenshtein distance, checking for all possible strings of forms on edit distance[3] 1 and 2, and 2) a neural network approach learning character-level sequences.[4] A corpus of ca. 100,000 tokens was compiled from 24 thematically related, mainly Early Irish texts published on Corpus of Electronic Texts (CELT).[5] The first version of the lemmatiser shows 45.2% accuracy[6] with unknown words and 71.6% with known words, while the neural network metrics are 64.9% and 99.2%, respectively; the neural model based on character-level sequences thus greatly outperforms the edit distance approach.

[3] Minimum edit distance, an approximate matching technique widely used in Natural Language Processing, measures how similar two strings are by calculating the minimum number of editing operations (insertion, deletion, substitution) needed to transform one string into another. In one of the most well-known variants, the Levenshtein distance, particular costs are assigned to each of these operations (Jurafsky and Martin, 2009, p.74).

[4] Code available at https://github.com/ancatmara/early-irish-lemmatizer [accessed 21 December 2020].

[5] https://www.ucc.ie/celt/ [accessed 19 December 2020].

[6] The percentage of correctly produced lemmas.

In the context of the ERC-funded project *Chronologicon Hibernicum* (ChronHib),[7] collections of Old Irish texts and glosses have been manually parsed using detailed morphological annotation and made available in digital database format as *Corpus Palaeohibernicum* or CorPH (Lash et al., 2020).

The Parsed Old and Middle Irish Corpus or POMIC (Lash, 2014b) consists of 14 manually POS-tagged and syntactically parsed texts from ca. 700–1100 A.D., totalling 33,000 words. The tagset is based on the one designed for the Penn-group of corpora for historical English. This corpus is currently being enlarged in the context of the ChronHib project. Finally, there are ongoing efforts to incorporate Early Irish material into the Universal Dependencies (UD) framework.[8]

3 The Old Irish Verb in a Nutshell

As McCone (1997, p.17) has pointed out, "[l]ike Modern Irish and Scots Gaelic, Old Irish is a basically verb-initial language in which the order verb-subject-object (VSO) predominates, except in the case of proclitic infixes or suffixes". However, as in Modern Irish, additional structures are found, especially with regard to the subject position in Old Irish (see, e.g., Mac Coisdealbha and Isaac, 1998 and Lash, 2014a).

The verbal complex (McCone, 1997, pp.1–19) comprises everything that falls within the accentual domain of the verb. This complex is highly synthetic with both fusional and agglutinative features (see example 1). Apart from the copula, no subject pronouns are used: person/number is encoded in the verb ending. The citation form of the verb in Early Irish is independent present indicative 3sg. Old Irish verb morphology is best described by lexical verb type and the opposition independent/dependent in the verbal complex. The interaction between these criteria translates into four possibilities, as shown in table 1.

Verb type	Independent	Dependent
Simple	Absolute *beirid* "carries"	Conjunct *ní·beir* "does not carry"
Compound	Conjunct (deuterotonic) *do·beir* "brings, gives"	Conjunct (prototonic) *ní·tabair* "does not bring, give"

Table 1: Interaction between Old Irish verb type and independent/dependent.

Apart from some tense/mood combinations, two ending sets exist. Absolute endings only occur with simple verbs, consisting of verb root plus ending (e.g., *beirid* "carries"). An invariably proclitic conjunct particle (e.g., *ní* "not") triggers a dependent form, and demands a conjunct ending with simplexes. Compound verbs, in contrast to simple verbs, are preceded by up to four lexical preverbs, e.g., *do·beir* "brings, gives" (with one preverb, namely **to*),[9] and invariably take (the same set of)

[7] https://www.maynoothuniversity.ie/chronologiconhibernicum [accessed 19 December 2020].

[8] For the currently available Old and Middle Irish tagged texts see https://universaldependencies.org/ [accessed 19 December 2020].

[9] Both *beirid* and *do·beir* consist of the verb root *ber* "carry".

conjunct endings.

Accentual patterns are integral to verb stem formation in Old Irish. By default, the stress is on the first syllable of the verbal complex, unless a proclitic conjunct particle or a preverb is present, in which case the stress shifts to the second "slot".[10] Since the first preverb in an independent compound verb is realised as a proclitic, this stem variant is called "deuterotonic", i.e., the stress is on the second element of the verb itself. In dependent position, the proclitic slot in the verbal complex is occupied by a conjunct particle, resulting in the first preverb of a compound verb to come under the stress (hence "prototonic"). Unlike deuterotonic forms, which have "a kind of barrier or juncture across which certain otherwise normal processes do not occur", prototonic forms have their first preverb "fully incorporated into the rest of the verb" (McCone, 1997, p.4). Compare, for example, deuterotonic *do·beir* with prototonic *·tabair* in table 1.

The verbal complex allows further infixation and suffixation of particles and pronouns. Object pronouns, mostly accusative, do not occur independently and are typically infixed, immediately preceding the proclitic juncture. Emphasising particles (Stifter, 2006, pp.127–128), or *notae augentes*, are invariably suffixed. The verb form *nondob·molor-sa* (Stokes and Strachan, 1901–1910, vol. i, p.593) in example (1) consists of the verb stem *mol-* with various affixes.[11] It contains, from left to right, the meaningless conjunct particle *no*, a consonant mutation affecting initial *d-* of the pronoun signalling a "nasalising relative clause" (*GOI* §§ 497–504) triggered by the temporal/causative conjunction *hóre*, an object pronoun 2pl., the verb stem, a deponent[12] present 1sg. ending, and a *nota augens* 1sg.

(1) (hóre) **no-n-dob-mol-or=sa**
(conjunction) CONJ_PART-REL-OBJ.2.PL-praise-PRS.IND.1SG=1SG
"(because) I praise ye" (Würzburg glosses 14c18)

The stress system of Old Irish often results in "complex synchronic morphophonemic alternations" (Stifter 2009: 90) and, consequently, a system of "double stem formation" (Russell, 2005, p.431). A key feature of the stress system is syncope, the deletion of unstressed vowels in even-numbered (but not final) syllables counting from the first stressed syllable onwards, with concomitant changes to the quality (i.e., (non-)palatalisation) of surrounding consonants, as well as to the form of verb endings. Compare *do·beram* in example (2) with dependent/prototonic *ní·taibrem* in example (3), the latter

[10] Note that the proclitic "slot" can take up more than one syllable; the stressed syllable of the verbal complex is therefore not necessarily the second one.

[11] Glossing is according to the Leipzig conventions for interlinear morpheme-by-morpheme glosses, for which see https://www.eva.mpg.de/lingua/resources/glossing-rules.php [accessed 19 December 2020]. Apart from the standard abbreviations, this paper includes the following abbreviations specified by the author: CONJ_PART = conjunct particle, PRT = preterite, PV = preverb.

[12] Apart from "normal" active endings, there are separate inflectional endings known as deponent, also conveying an active meaning, constituting a "merely lexical property that has to be known for each verb separately" (Stifter, 2009, p.87).

showing the deletion of the root vowel *e* in *ber* and subsequent phonological (and orthographic) processes. Note also that the preverb **to* is fully incorporated into the verb stem in *ní·taibrem*, in contrast to *do·beram*, where it appears as (unstressed) *do*.

(2) **do-ber-am**
PV-carry-PRS.IND.1PL
"we bring, give"

(3) **ní-taibr-em**
CONJ_PART.NEG-bring-PRS.IND.1PL
"we do not bring, give"

Stem allomorphy and syncope constitute key aspects of the finite-state implementation (see section 5.2). In general, the Old Irish verbal system most clearly shows—out of all the grammatical subsystems of Old Irish—how phonology imposes itself to a great extent upon the morphology.

4 Finite-State Morphology

Morphological parsing of Old Irish verb forms has been implemented using the paradigm of finite-state morphology (Beesley and Karttunen, 2003). State machines, or automata, recognise a particular set of sequences of symbols (strings) as defined by a regular expression. Automata can be conceptualised as networks with transitions through a finite number of paths. Finite-state transducers (FSTs) are finite-state automata with two-level relations for each path in the network. These inherently bidirectional mappings are very well suited for linguistic modelling, especially morphology, employing the notion of a lexical and surface level.

Figure 1 visualises a finite-state transducer as a network for the orthographic string *léicid* (prs. ind. 3sg., "lets"). The term analysis is used for the process whereby a lexical-level (commonly upper-level) string is produced from a surface form; this is effectively morphological parsing. The opposite mapping process is called generation, resulting in a surface-level (or lower-level) string. Each path represents a one-symbol transition (a multicharacter string such as `+VROOT` is also one symbol). The symbol ε (epsilon) denotes an empty transition and translates into a mapping with no accompanying symbol on the opposite level, i.e., when the upper-level and lower-level strings are of unequal length.

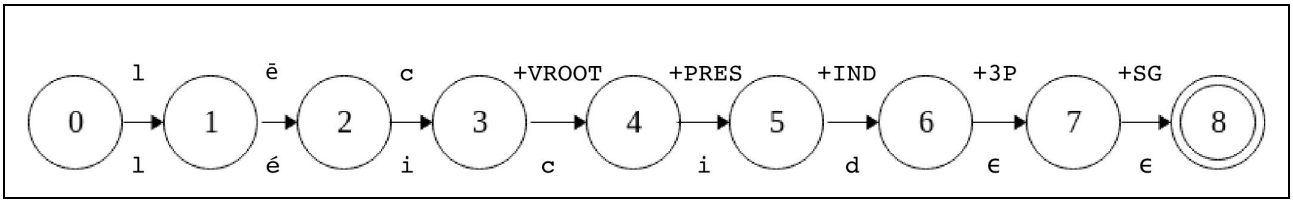

Figure 1: A finite-state transducer (FST) accepting, at final state 8, a set of two-level symbol mappings:

`l ē c +VROOT +PRES +IND +3P +SG : l é i c i d`

Various finite-state toolkits are available. Xerox Finite State Tool (*xfst*) is the one accompanying

Beesley and Karttunen (2003).[13] It incorporates an extended set of regular expression operators to model the morphotactics and morphophonemic processes (conditional rewrite rules) of a language. This toolkit includes a Lexicon Compiler (*lexc*) (Chapter 4 in Beesley and Karttunen, 2003), which aids the modelling of the morphotactics of a language and simplifies the creation of a natural-language lexicon. The open-source toolkit *foma*[14] (Hulden, 2009) is compatible with *xfst* and *lexc* and has been used by the author to create a finite-state morphological transducer for Old Irish.

5 Implementation

The finite-state morphological transducer for Old Irish is available online.[15] Two challenges in encoding morphological and orthographic features using finite-state transducers will be discussed in this section: word segmentation and stem allomorphy.

5.1 Word Segmentation

5.1.1 Spacing and Morpheme Boundaries

Modelling the Old Irish verbal complex entails catering for varying word segmentation practices. Different editorial standards result in the use of different kinds of typographical markers between proclitic element(s) and the stressed verb stem. With elements such as the negative particle *ní*, we often find spacing in older text editions. Modern text editions and grammars employ the raised dot "·". However, with diplomatic editions, i.e., when the transcribed text is faithful to the manuscript, no explicit boundary markers are present; the medieval scribe may or may not have employed a space. We thus may encounter example (4) as *níléici*, *ní·léici* or *ní léici*, depending on the editorial policy relative to the text in which the form is found.

(4) **ní-léic-i**
CONJ_PART.NEG-let-PRS.IND.3SG
"does not let"

This variability leads to two complementary challenges. The absence of a space (or a segmentation marker) necessitates the encoding of dependencies on the level of the entire verbal complex. For example, when a verb follows the conjunct particle *ní*, it can only have a conjunct ending. If this conjunct particle is immediately consecutive to *léici*, the automatic parse should only include a conjunct ending reading, such as prs. ind. 2sg. or 3sg. Spacing, however, causes morphotactic dependencies to transcend what we could now call the word boundary. If we encounter *léici* "on its own", and since automatic morphological parsing is word based, we should also incorporate the possibility that this is an absolute prs. ind. 2sg. form (obviously assuming that no conjunct particle is present), which happens to be identical to the conjunct 2sg. and 3sg. forms. This paper reports on the creation of a morphological FST, the task of which is to present all possible word-level analyses; the subsequent step of word-transcending disambiguation, whether due to editorial spacing policy or not,

[13] http://www.fsmbook.com [accessed 19 December 2020].
[14] https://fomafst.github.io/ [accessed 19 December 2020].
[15] https://github.com/ThFransen84/OIfst [accessed 19 December 2020].

is not part of the initial task of morphological parsing.

With non-space boundary markers such as hyphens we have a choice: either delete them in the pre-processing stage, or leave them in. In the current implementation phase the boundary markers "-" and "·" are (optionally) incorporated as boundary marker. This means that forms with a boundary marker are being analysed as one string, and that dependencies/restrictions across proclitic elements and verb roots/stems must be encoded—if the goal, of course, is to exclude, as much as possible, morphotactically illegal strings in the morphological transducer.

Automatic tokenisation for Old Irish has only very recently started to receive attention. Doyle et al. (2019) report on the development of neural machine-learning methods for tokenising the Old Irish Würzburg glosses. It is hoped that future collaboration will generate advances and solutions for uniform segmentation practices and word-level parsing for Old Irish.

5.2 Encoding Morphotactic Dependencies

The challenges of spacing and morphotactic dependencies have been tackled by creating separate, yet combinable, transducers for proclitic elements ("prefixes") and the verb root stem proper (including endings), respectively. Dependencies have been encoded by using two morphotactic restriction methods available in the finite-state paradigm:

- Flag Diacritics (Chapter 7 in Beesley and Karttunen, 2003). Symbols that can be inserted alongside morphemes in the concatenation architecture to control which paths are allowed and which should be blocked in the network. Flag Diacritics do not interfere with the process of inputting (analysing) or outputting (generating) a string and can be made invisible in the output. Furthermore, they may be deleted from the network, removing illegal paths but leaving legal paths intact. One needs to carefully think about separated dependencies in advance when using Flag Diacritics, since "adding Flag Diacritics *post hoc* to an existing system can require non-trivial re-editing of your source files" (Beesley and Karttunen, 2003, p.340).
- Upper-level filters (Beesley and Karttunen, 2003, pp.249–254). This method involves the creation of lexical-level tag combinations defined as the complement language, i.e., the strings that should not be part of the language, which are subsequently deleted from the network. In contrast to Flag Diacritics, these operations apply *after* the creation of the transducer, initially allowing for overgeneration.

Flag Diacritics were found to be very convenient for modelling the (often long-distance) dependencies regarding compound verbs, where the combination of preverb and stem is arbitrary (the various preverbs do not go with every verb stem, and many verb stems cannot be preceded by a preverb). Other dependencies involve more complex restriction specifications for which the *post-hoc* upper-level filters were found to be more suited.

5.3 Dealing with Stem Allomorphy

As mentioned in section 3, due to the impact that phonology has on Old Irish morphology, Old Irish exhibits a significant degree of allomorphic variation in verb stem formation. A possible solution to

cater for non-trivial and synchronically often highly unpredictable stem formation is operating with what is called a "monolithic stem" in the author's implementation: a base form potentially consisting of multiple morphemes from a historical perspective, but encoded on the surface level as a synchronic, more or less invariant stem alternant.

The base forms to be encoded correlate strongly with the system of alternating stems which are the result of either the verb root or a preverb being in stressed position. By assigning historical roots to the lexical/upper level in the finite-state morphological architecture, we can create a mapping with stems on the surface/lower level as illustrated with *do·sluindi* "denies" below. Example (5) is the independent, deuterotonic form and example (6) is the prototonic form. Verb stems and their accompanying historical roots are marked with a border.

(5) d ī +PV + s l o n d +VROOT : d o · s l u i n d

(6) d ī +PV + s l o n d +VROOT : d í l t

Since syncope (internal vowel deletion) is mechanically applied by a conditional rewrite rule in the author's finite-state rule framework, [16] stem entries are pre-syncope, i.e., showing underlying internal vowels (which often do not surface at all). Consequently, stem encoding entails that for each verb (whether simple or compound) a list of pre-syncope and hence semi-surface-level stems needs to be specified.

The notion of stem discussed in this section is different from the more current usage of stem in Old Irish grammars, which talk about a present, subjunctive, future, preterite active and preterite passive stem. There are six groups of inflectional ending sets which are not arbitrarily combinable with each of these five stems (see Stifter, 2009, p.88 for the combinations). However, the relevant ending set is associated with stem type, and in the author's implementation each of the monolithic stems already incorporates the verb class and relevant stem formation process. In other words, all we need to do is specify the correct ending set for each monolithic stem entry in *lexc* (see section 4).

The number of necessary monolithic stems depends on whether the verb is simple or compound, weak or strong, and regular or irregular (containing suppletive stems in its paradigm).[17] Since (tense/mood) stem formation with weak verbs is both (mostly) predictable and transparent relative to the root, we only need between one and four monolithic stems for verbs belonging to this class. The four monolithic stems for *do·léici* "lets go, casts" are found in example (7). As said above, these stems are pre-syncope and can be viewed as semi-surface bases. The last two bases contain what is called the

[16] The application of syncope is often counteracted, both by predictable and unpredictable processes. Irregular application of syncope (see Ó Crualaoich, 1999 and also the examples in *GOI* § 109) is hard to accommodate in a rule-based approach. The exact details relating to (the computational implementation of) syncope are rather technical; readers interested in a comprehensive discussion of this topic are referred to Fransen (2019, pp.21–22 and pp.82–85).

[17] Ignoring the small group of hiatus verbs with roots ending in a vowel.

augment in Early Irish grammar (McCone, 1997, pp.127–161; Stifter, 2006, pp.250–256), a particle most commonly appearing as *ro*, supplying resultative or potential meaning and formally behaving like a preverb.

(7) *do·léic-*
·teilic-
do·reilic-
·tarolaic-

Strong verbs exhibit stem changes across their tense/mood paradigms which only become transparent when one knows the abstract root shape. Strong verbs realistically demand at least five base entries for each of the tense/mood stems. For strong types of compound verbs, with diverging deuterotonic and prototonic forms, we easily reach twice this amount when formulating monolithic stems.

The author's focus has been on weak verbs since stem formation can be straightforwardly automated with this class of verbs. Having said this, the architecture of the *lexc* file facilitates integration of strong verbs and irregular (suppletive) stems. It is the conviction of the author that the stem-encoding approach described in this section is the only feasible one in order to arrive at a synchronically oriented, automatic rule-based morphological framework that can both correctly generate and analyse Old Irish verbs.

6 Results

As discussed in section 5.2, the focus in the finite-state implementation is on weak verbs, the stem formation of which is (mostly) predictable and transparent. After implementation of weak verb inflection, the transducer was tested on the (partly reconstructed) Old Irish text *Táin Bó Fraích* "The cattle-raid of Fróech" (Meid, 1974). The edited text is available on CELT[18] and was subject to earlier computational investigations by Lynn (2012). The text consists of a total of 50 Old Irish weak verb forms (types), excluding verbal nouns, categorised under 27 lemmas; 36 out of 50 inflected forms (72%) were recognised. Table 2 shows an excerpt of the results; for a complete overview see Fransen (2019, p.111).

[18] https://celt.ucc.ie//published/G301006/ [accessed 19 December 2020]. The digitised edition does not contain the vocabulary section in the print edition, i.e., the index of lemmas with accompanying inflected forms.

Lemma	Form	Morphological segmentation	Morphological gloss	Analysed correctly?
ad·ella	*aidleth*	aidl-eth	approach-IPFV.3SG	Yes
brissid	*brissis*	briss-is	break-PRT.3SG	Yes
do·léici	*Dolléici*	do-lléic-i	PV-let-PRS.IND.3SG	Yes
	Dolléicther	do-lléic-ther	PV-let-PRS.IND.PASS.3SG	Yes
	Dolléicetar	do-lléic-etar	PV-let-PRS.IND.PASS.3PL	Yes
fo·dáili	*Fodlid*	fodl-id	distribute-IMP.2PL	No
marbaid	*marbam*	marb-am	kill-IMP.1PL	Yes

Table 2: Results for seven verbs across five lemmas.

7 Discussion

Old Irish orthography may result in grammatical ambiguity. For example, the Old Irish spelling system is only unambiguous with nasalised <b, d, g>, which appear as <mb>, <nd> and <ng>, respectively, and lenited <c, p, t>, which are rendered as <ch, ph, th>, respectively (see Stifter, 2006, pp.377–378) for an overview of phoneme-to-letter correspondences). All forms with lemma *do·léici* in table 2 were originally wrongly parsed as relative forms, since the doubling of some consonants, including <ll>, was implemented only as a diagnostic for a nasalising relative clause (*GOI* §§ 497–504; see also example 1). The complicating factor is that doubling of consonants, such as with the digraph <ll> in this case, is alternatively employed by scribes to mark that the consonant is unlenited (*GOI* § 136). Upon closer inspection, it was found that all forms of *do·léici* are sentence-initial (starting with a capital letter in the text edition) and therefore cannot introduce a nasalising relative clause. This problem was fixed by encoding <ll> in *anlaut* position as an orthographic variant of unlenited <l> and introducing a capitalisation rule to rule out the relative reading.

The finite-state morphological transducer reported on in this paper aims to analyse and generate mainly Classical Old Irish forms and spelling, adhering to an arguably somewhat superficial orthographic standard. The inflectional ending *-id* such as in *fodlid* (see table 2), for example, is invariably encoded in the transducer as *-aid* (e.g., *fodlaid*), which is not ambiguous as to the quality of the preceding consonant or consonant cluster (hypothetically, *fodlid* could also represent *foidlid*). Ubiquitous, minor and transparent variation of this type is probably handled best by creating a separate and subsequent "spelling transducer", which maps a canonical spelling or phonological representation to a (finite) number of character variants, e.g., the set of possible written vowels or vowel combinations surrounding palatal and non-palatal consonants. This remains future work.

For more extensive variation, or spellings that are further removed from an Old Irish standard (e.g. Middle Irish forms), we can resort to one of the lemmatiser implementations described in Dereza (2018) (see section 2), using algorithms to predict a lemma on the basis of either string similarity measures or neural networks using character-level sequences, based on known inflected forms (the latter, however, are indiscriminately Old, Middle or Early Modern Irish, as they have been taken, without explicit chronological information, from eDIL). By augmenting the lemmatiser dictionary with inflected forms generated by the author's Old Irish morphological transducer, we would create

a normalisation component that better predicts "proper" Old Irish spellings for synchronic and diachronic variants, as well as significantly increasing the power of the lemmatiser itself.

A best practice in relation to treatment/encoding of typographical markers indicating morpheme boundaries (see section 5.1.1), as well as subsequent morphosyntactic disambiguation strategies, will hopefully develop in collaboration with key players in the field of Old Irish computational linguistics.

8 Conclusion

The present paper very much reflects work in progress. The finite-state implementation, especially in relation to tackling Old Irish verb stem allomorphy, is promising as well as linguistically interesting. Admittedly, adding to the inventory of stems and rules is manually expensive and relies on expert knowledge. The potential advantage of supplementary machine learning methods may be worth investigating, especially in relation to unpredictable grammatical variation and irregularity (e.g., unexpected syncope patterns and paradigm levelling).

Future plans include building transducers for the remaining parts-of-speech and working towards the creation of a POS-tagger for Old Irish, which can be adapted to cover later periods, notably Middle Irish. Ways of collaboration between the projects mentioned in section 2 and the author's one have been, and continue to be, investigated. The author's ultimate goal is to link up linguistic resources for Old and Modern Irish and seal the "lexicographical gap" that currently exists between these periods.

9 References

Beesley, K.R. and Karttunen, L. 2003. *Finite-state morphology.* Stanford, CA: Center for the Study of Language and Information (CSLI) Publications.

Dereza, O. 2018. Lemmatization for ancient languages: rules or neural networks?. In: Ustalov, D., Filchenkov, A., Pivovarova, L. and Žižka, J. eds. *Artificial Intelligence and Natural Language: 7th International Conference, AINL 2018, 17–19 October 2018, St. Petersburg*. Cham: Springer, pp.35–47.

Doyle, A., McCrae, J.P. and Downey, C. 2019. A Character-Level LSTM Network Model for Tokenizing the Old Irish text of the Würzburg Glosses on the Pauline Epistles. In: *Proceedings of the Celtic Language Technology Workshop 2019, 19 August 2019*, *Dublin, Ireland.* [Online]. European Association for Machine Translation, pp.70–79. [Accessed 19 December 2020]. Available from: https://www.aclweb.org/anthology/W19-6910/

Fomin, M., and Toner, G. 2006. Digitizing a Dictionary of Medieval Irish: the eDIL project. *Literary and linguistic computing*. **21**(1), pp.83–90.

Fransen, T. 2019. *Past, present and future: Computational approaches to mapping historical Irish cognate verb forms*. Ph.D. thesis, Trinity College Dublin, The University of Dublin.

Fransen, T. 2020. Automatic morphological analysis and interlinking of historical Irish cognate verb forms. In: Lash, E., Qiu, F. and Stifter, D. eds. *Morphosyntactic Variation in Medieval Celtic Languages*. Berlin, Boston: De Gruyter Mouton, pp.49–84.

Hulden, M. 2009. Foma: a finite-state compiler and library. In: *EACL '09: Proceedings of the 12th*

Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session, 3 April 2009, Athens. [Online]. Association for Computational Linguistics, pp.29–32. [Accessed 19 December 2020]. Available from: https://dl.acm.org/doi/10.5555/1609049.1609057

Jurafsky, D. and Martin, J.H. 2009. *Speech and Language Processing: an introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* 2nd ed. Upper Saddle River, NJ: Prentice Hall.

Lash, E. 2014a. Subject positions in Old and Middle Irish. *Lingua*. **148**, pp.278–308.

Lash, E. 2014b. *The Parsed Old and Middle Irish Corpus (POMIC)* (Version 0.1). [Software]. [Accessed 19 December 2020]. Available from: https://www.dias.ie/celt/celt-publications-2/celt-the-parsed-old-and-middle-irish-corpus-pomic/

Lash, E., Qiu, F. and Stifter, D. 2020. Introduction: Celtic Studies and Corpus Linguistics. In: Lash, E., Qiu, F. and Stifter, D. eds. *Morphosyntactic Variation in Medieval Celtic Languages*. Berlin, Boston: De Gruyter Mouton, pp.1–12.

Lynn, T. 2012. Medieval Irish and Computational Linguistics. *Australian Celtic Journal*. **10**, pp.13–27.

Mac Coisdealbha, P. and Isaac, G.R. ed. 1998. *The syntax of the sentence in Old Irish: selected studies from a descriptive, historical and comparative point of view*. Tübingen: Niemeyer.

McCone, K. 1985. The Würzburg and Milan Glosses: our earliest sources of "Middle Irish". *Ériu*. **36**, pp.85–106.

McCone, K. 1997. *The Early Irish verb*. Revised edition with *index verborum*. Maynooth: An Sagart.

Meid, W. ed. 1974. *Táin Bó Fraích*. 2nd ed. Dublin: Dublin Institute for Advanced Studies.

Ó Cróinín, D. 2001. The earliest Old Irish glosses. In: Bergmann, R., Glaser, E. and Moulin-Fankhänel, C. eds. *Mittelalterliche volkssprachige Glossen: Internationale Fachkonferenz des Zentrums für Mittelalterstudien der Otto-Friedrich-Universität Bamberg 2. bis 4. August 1999*. Heidelberg: Winter, pp.7–31.

Ó Crualaoich, C. 1999. *Some irregular syncope patterns in Old Irish*. Ph.D. thesis, National University of Ireland, Maynooth.

Ó Muircheartaigh, P. 2015. *Gaelic dialects past and present: a study of modern and medieval dialect relationships in the Gaelic languages*. Ph.D. thesis, University of Edinburgh, UK.

Piotrowski, M. 2012. *Natural Language Processing for historical texts.* San Rafael: Morgan & Claypool Publishers.

Russell, P. 2005. What was best of every language: the early history of the Irish language. In: Ó Cróinín, D. ed. *Prehistoric and early Ireland, vol. 1*. Oxford: Oxford University Press, pp.405–450.

Stifter, D. 2006. *Sengoídelc: Old Irish for beginners.* Syracuse, New York: Syracuse University Press.

Stifter, D. 2009. Early Irish. In: Ball, M.J. and Müller, N. eds. *The Celtic Languages*. 2nd ed. Abingdon and New York: Routledge, pp.55–116.

Stokes, W. and Strachan, J. eds. 1901–1910. *Thesaurus palaeohibernicus: a collection of Old-Irish glosses, scholia, prose, and verse*. 3 vols. Cambridge: Cambridge University Press.

Thurneysen, R. 1946. *A Grammar of Old Irish.* Dublin: Dublin Institute for Advanced Studies.

Toner, G., Ní Mhaonaigh, M., Arbuthnot, S., Theuerkauf, M. and Wodtko, D. 2019. *Electronic Dictionary of the Irish Language* (www.dil.ie 2019). [Online]. [Accessed 19 December 2020].

Available from: http://dil.ie/

Uí Dhonnchadha, E., Scannell, K., Ó hUiginn, R., Ní Mhearraí, E., Nic Mhaoláin, M., Ó Raghallaigh, B., Toner, G., Mac Mathúna, S., D'Auria, D., Ní Ghallchobhair, E. and O'Leary, N. 2014. Corpas na Gaeilge (1882–1926): Integrating Historical and Modern Irish Text. In: Bjarnadóttir, K., Driscoll, M., Krauwer, S., Piperidis, S., Vertan, C. and Wynne, M. eds. *Proceedings of the LREC 2014 Workshop LRT4HDA: Language resources and technologies for processing and linking historical documents and archives*, *26 May 2014, Reykjavik*. [Online]. European Language Resources Association (ELRA), pp.12–18. [Accessed 19 december 2020]. Available from: http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-LRT4HDA%20Proceedings.pdf

Acknowledgements

This paper is based on research carried out during a Government of Ireland Postgraduate Scholarship (GOIPG/2017/1808) funded by the Irish Research Council. I would also like to acknowledge the two anonymous reviewers for providing helpful comments on this paper. All remaining errors and omissions are, of course, my own.