Theodorus Fransen

3 Automatic morphological analysis and interlinking of historical Irish cognate verb forms

1 Introduction

The main aim of the author's research project is to use computational approaches to gain more insight into the historical development of Irish verbs. One of the objectives is to investigate how a link between the electronic Dictionary of the Irish language (eDIL),[1] covering the period c. 700–c. 1700, but focussing on Early Irish (7th–12th centuries), and the nascent *Foclóir Stairiúil na Gaeilge* 'The Historical Dictionary of Irish',[2] covering the period 1600–2000, could be implemented. Such a link will be hugely beneficial for scholars operating at the intersection of the medieval and modern period (see Table 1), who currently lack a comprehensive lexical resource for the "intermediate" early modern period.

The above-mentioned lexicographical discontinuity is problematic, and needs to be remedied, especially in the light of the pervasive changes in the verbal system between Early and Modern Irish. The author's motivation for focussing on the verbal system in Early Irish resonates with the following observation by McCone in his authoritative monograph on the Early Irish verb:

> Concentration upon the verb was dictated by its generally conceded status as the most difficult and interesting area of Old and Middle Irish morphology and few would deny that an understanding of the Old Irish system's workings and development into and through Middle Irish is a prerequisite for being able to deal with the abundance of Old and Middle Irish texts effectively. (McCone 1997: xviii)

During the author's research it was found that eDIL does not provide full verb paradigms for many verb entries. It was felt that additional language technology is necessary to deal with the complex Early Irish verbal system. Such technology will also facilitate more systematic and comprehensive interlinking of verb forms in lexicographical resources. The main contribution to this end by the author is the development of a morphological analyser for Old Irish, which is also the focus of this paper.

1 Available at: http://dil.ie/ [accessed 7 February 2019].

2 Available at: https://www.ria.ie/research-projects/focloir-stairiuil-na-gaeilge [accessed 7 February 2019].

https://doi.org/10.1515/9783110680744-004

In order to make this contribution accessible to (computational) linguists whose research area is not Old Irish, a brief overview of the Irish language periods (section 2) and the basics of the Old Irish verbal system (section 3) is provided. The latter aims to show how phonology imposes itself on verb morphology, resulting in an often complex relationship between an underlying verb root and a verb's multiple surface shapes – an insight crucial for the computational implementation. Section 4 sums up important changes in the verbal system in Middle Irish and beyond. In the second half of the paper, the focus is on digital resources for historical Irish and Natural Language Processing methods. Section 5 surveys important existing digital resources and computational methods used to deal with historical texts. The proposed methodological framework of the paper is the topic of section 6. Section 7 introduces finite-state morphology and presents some highlights, as well as challenges, in the development of a morphological analyser for Old Irish verbs. The formulation of clear-cut verb stem entities constitutes a key feature in the implementation. Suggestions for automatically linking cognate verb forms are presented in section 8. A synthesis of matters discussed in this paper follows in section 9, which also outlines some research prospects.

2 A historical sketch of the Irish language

The historical period of Irish can be divided into the language stages shown in Table 1 below. Greene (1966) provides a succinct overview of the history of the Irish language. Early Irish represents the language from the early medieval period up until about 1200. After that we speak of Modern Irish. Old Irish, like the modern standardised language, can be treated as a normative phase in the

Table 1: Medieval and Modern Irish language periods.

	Language stage	Time period
Early Irish	Old Irish	7th–9th centuries A.D.
	Middle Irish	10th–12th centuries
Modern Irish	Early Modern Irish (including Classical Modern Irish)	13th –mid-17th centuries
	Post-Classical Modern Irish	mid-17th–mid-19th centuries
	Irish of the Revival period	late 19th–early 20th centuries
	contemporary standardised Modern Irish	1958–present

history of the language. Indeed, "classical" Old Irish, the language as witnessed predominantly in Old Irish glosses in Latin manuscripts, is the basis for many grammars and handbooks, including *A Grammar of Old Irish* by Rudolf Thurneysen (1946) (GOI). While representing a stable and normative phase in the language's history (McCone 1997: 166), Old Irish shows diachronic as well as synchronic linguistic variation (for a discussion of the latter see McCone 1985).[3] However, the linguistic variation in Old Irish is negligible compared to the unstable and highly variable language seen in Middle Irish texts. As McCone (1997: 166–167) has pointed out, Middle Irish comprises standard Old Irish forms and forms anticipating Modern Irish usage, as well as forms that are consonant with neither. The end of the Middle Irish period sees the production of the great medieval Irish manuscripts.[4]

The subsequent Early Modern Irish period (13th–mid-17th centuries) is dominated by a literary genre of praise poetry in syllabic verse composed by court poets, referred to as Classical Modern Irish (McManus 1994). In contrast to the highly regulated grammar of this bardic poetry, however, hugely varying registers can be observed with prose texts of this period, ranging from archaic language to registers that are not far removed from 19th-century Irish (Ó hUiginn 2013: 87–89).

Post-Classical Modern Irish refers to the literary period between the downfall of the Irish-speaking aristocracy in the early 17th century and the Great Famine (1845–1849), which is characterised by – amongst other developments – a more regional orientation in writing (Ó Háinle 2006). The classical literary standard that had emerged in the early modern period gradually gives way to writing conventions that more closely reflect the contemporary spoken language, resulting in the coming to the fore of the Irish dialects in texts of this period (Williams 1994).

The period between the Great Famine and the creation of the Free State (1922) is known as the Gaelic Revival, which witnessed an increased production of original work, facilitated by institutions such as the *Conradh na Gaeilge* [Gaelic League], established in 1893 (Mahon 2006). After independence, plans were made for a standardisation of Irish grammar and spelling, ultimately codified in a 1958 booklet published by the Irish government's Translation Department (with further revisions in 2012 and 2016).[5]

3 Ó Cróinín (2001) discusses diachronic orthographical developments in the earliest Old Irish glosses. Two conventions, the "Irish" and "British" system, seem to have competed with each other; the latter ultimately became the standard for all subsequent Irish literature.

4 *Lebor na hUidre*, Rawlinson B 502 and The Book of Leinster (*An Leabhar Laighneach*) (Breatnach 1994a: 222–225).

5 Available at: https://data.oireachtas.ie/ie/oireachtas/caighdeanOifigiul/2017/2017-08-03_an-caighdean-oifigiuil-2017_en.pdf [accessed 7 February 2019].

3 The Old Irish verb: The morphology-phonology interface

3.1 The main skeleton of the verbal complex

In general, Old Irish is a VSO language (Russell 2005: 430). However, additional variant structures are found (Mac Coisdealbha 1998), especially regarding the subject position (Lash 2014b). Both the verb, subject and object may be contained within the "verbal complex" (see McCone 1997: 1–19), comprising everything that falls within the accentual domain of the verb (Stifter 2009: 84), as such potentially constituting a highly synthetic "word". Leaving aside copula constructions, Old Irish inflected verb forms incorporate the subject; no independent subject pronouns exist. Third person forms – from the viewpoint of word-based parsing – are inherently ambiguous in that there might or might not be an independent subject. In the present work, as is customary, third person verb forms are not glossed with a pronominal subject in the English translation.

There is a distinction between "simple" and "compound" verbs. Verbs with the verb root as their sole lexical element, as in (1) and (2), containing root *ber*, are called simple. A compound verb additionally takes one or more preceding lexical preverbs (PV), originating in prepositions,[6] modifying the meaning of the verb root. In (3) and (4), the preverb is underlying/historical *to* combined with the root *ber*. As a rule, the first preverbal element within the accentual domain of the verb is realised as a proclitic, resulting in a juncture between, put simply, a prefix and the stressed part of the verbal complex, as in (2)–(4). This juncture is denoted by a mid-high dot to facilitate grammatical analysis; it is not present in manuscripts.

(1) *beirid*
carry$_{3SG.PRES}$
'carries'

(2) *ní·beir*
NEG·carry$_{3SG.PRES}$
'does not carry'

(3) *do·beir* (*to-ber-*)
PV·bring$_{3SG.PRES}$
'brings'

6 Called thus by Thurneysen (GOI §§ 819–821).

(4) *ní·tabair* (*to-ber-*)
NEG·bring$_{3SG.PRES}$
'does not bring'

Some commonly used grammatical notions relating to stem and ending formation are key to understanding the workings of the verbal complex. First, there are two ending sets, "absolute" and "conjunct". Only simple verbs can take absolute endings, and only when occurring in clause-initial position. The conjunct ending set applies when a verb is conjoined with a preverbal element; compound verbs therefore invariably carry conjunct endings, while simple verbs take this set of endings when preceded by the preverbal "conjunct particles" (C), e.g. the negative particle *ní* 'not', as in (2). In (1), *-id* is the third singular present indicative absolute ending. The corresponding conjunct ending is seen in (2)–(4), where palatalisation of the root-final *r* (orthographically encoded by preceding *i*) is the only marker of inflection.

A verb preceded by a conjunct particle is said to be dependent, and independent otherwise. The distinction between independent and dependent has major repercussions for the surface shape of especially compound verbs. Generally speaking, an independent compound verb appears in its "deuterotonic" form as the first preverb is realised as a proclitic, causing the stress to fall on the second element (the verb root in [3]). When the proclitic "slot" in the verbal complex is occupied by a conjunct particle, as in the (dependent) compound form in (4), the stress is on the verb's first preverb; this stem alternant is accordingly called "prototonic".[7] As (3) and (4) illustrate, the stress system of Old Irish may result in "complex synchronic morphophonemic alternations" (Stifter 2009: 90) and, consequently, a system of "double stem formation" (Russell 2005: 431). The abundant allomorphic variation seen in the Old Irish verbal system raises a question crucial for implementational purposes: what exactly is a verb stem in Old Irish? Section 7.2.2 will detail how this question has been tackled from a computational point of view.

7 There are some exceptions to the rules laid out here in relation to dependency, the distinction deuterotonic/prototonic and the set of endings that is demanded. In the imperative only conjunct endings exist, and compounds in the imperative appear in their prototonic form regardless of dependency. A further anomaly exists with compounds whose first preverb is either *to, fo* or *ro*, which may equally assume their prototonic form in independent position if the following element starts with a vowel, causing vowel elision (McCone 1997: 3). Proclitic conjunctions such as *co* 'that' (GOI § 896) may be found with either independent or dependent verb forms, i.e. they sometimes assume the status of conjunct particle.

Compound verbs may take up to four preverbs, each of which adhere to a positional hierarchy tentatively formulated by McCone (1997: 89–90). Verb roots cannot be arbitrarily compounded with any preverb. However, most verbs are liable to being (further) compounded with an "augment". While a lexical preverb in origin, the augment has developed a "modificatory function that belongs to the grammar of Old Irish and not to its lexicon" (McCone 1997: 91). This preverbal particle supplies either a resultative or potential meaning, depending on the tense and/or mood of the verb form that it occurs with, illustrated with (5) and (6), respectively.

(5) *ro·léic*
AUG·let$_{3SG.PRET}$
'has let'

(6) *as·robair* (*ess-ro-ber-*)
PV·say$_{AUG.3SG.PRES}$
'can say'

The augment is most commonly *ro* (position 4), while the augments *ad* (position 3) and *cum* (position 4) are more restricted – i.e. the latter two co-occur with a limited set of (lexical) preverbs.[8] For a discussion on the preverbal particle *ro* and other augmentation strategies see GOI (§§ 526–537). Simple verbs (which do not have a preverb) almost always take *ro*, rather than *ad* or *cum*. The augment adds to the already abundant allomorphic variation seen in stem formation and its position is subject to change during the Early Irish period, in parallel with other processes of reorganisation and simplification of the verbal system,[9] most importantly the univerbation[10] of old compound verbs.

The morphosyntax or morphotactics of the verbal complex, i.e. the legal combination of morphemes (Beesley and Karttunen 2003: 26–27), with optional morphemes in brackets, is schematically summarised in (7) (C = conjunct particle, * = zero or more, with the provision that the total of preverbs does not exceed four, E = ending). Table 2 shows the schematic structure of the verbal complex with examples of preterite formations (with the conjunct third person singular

8 *ad* and *cum* are underlying forms, subject to a substantial amount of allomorphic variation depending on whether they are stressed or not.

9 Already in the Old Irish period, and during the Middle Irish period, *ro* is gradually adopting the status of conjunct particle, mitigating its effects in stressed position. The positional behaviour of *ro* and its semantics is outside the scope of the present paper; the reader should refer to McCone (1997: 127–161) for a detailed description of this preverbal particle.

10 A lexicalisation process involving the "unification . . . of a syntactic phrase or construction into a single word" (Brinton and Traugott 2005: 48).

Table 2: Schematic structure, including the position of the augment *ro*, of the Old Irish verbal complex, adapted from McCone (1997), with third person singular examples of unaugmented and augmented preterite forms with root *lēc*, illustrating combinatorial possibilities and allomorphic variation in stem formation (C = conjunct particle, * = zero or more, with the provision that the total of preverbs is not more than four). For Old Irish phonemes and their graphemic representation see Stifter (2006: 377–379).

Lemma		Structure (RO = augment *ro*)	Dependency	Ending	Example, pret. 3sg. (bold = lexical element, italics = stressed syllable)
léicid (simplex)	i.	VROOT E	indep.	abs.	***léic***is /ˈlʲeːgʲəsʲ/ $\text{let}_{\text{3SG.PRET}}$
	ii.	C · VROOT E	depend.		ní·***léic*** /nʲiː ˈlʲeːgʲ/ $\text{NEG·let}_{\text{3SG.PRET}}$
	iii.	RO · VROOT E	indep.		ro·***léic*** /ro ˈlʲeːgʲ/ $\text{AUG·let}_{\text{3SG.PRET}}$
	iv.	C · RO VROOT E	depend.		ní·*rei*lic /nʲiː ˈrʲelʲəgʲ/ (ro-lēc-) $\text{NEG·let}_{\text{AUG.3SG.PRET}}$
do·léici (compound)	v.	PV1 · PV* VROOT E	indep. (deut.)	conj.	**do·*léic*** /do ˈlʲeːgʲ/ (to-lēc-) $\text{PV·cast}_{\text{3SG.PRET}}$
	vi.	C · PV1 PV* VROOT E	depend. (protot.)		ní·*tei*lic /nʲiː ˈtʲelʲəgʲ/ (to-lēc-) $\text{NEG·cast}_{\text{3SG.PRET}}$
	vii.	PV1 · PV* RO PV* VROOT E	indep. (deut.)		**do·***rei***lic** /do ˈrʲelʲəgʲ/ (to-ro-lēc-) $\text{PV·cast}_{\text{AUG.3SG.PRET}}$
	viii.	C · PV1 PV* RO PV* VROOT E	depend. (protot.)		ní·*tar*laic /nʲiː ˈtarləgʲ/ (to-ro-lēc-) $\text{NEG·cast}_{\text{AUG.3SG.PRET}}$

form having a zero ending); it illustrates how the stress pattern (phonology) of Old Irish impacts on the verb morphology. The situation is slightly simplified in that *ro* represents the augment; *ro* is the particle's most common allomorph and the one found with *léicid* 'lets' and *do·léici* 'lets go, releases, casts'.

(7) (C) PV* (AUG) PV* VROOT E

3.2 Adjuncts and *notae augentes*

The skeleton of the verbal complex, outlined in (7), allows for incorporation of unstressed, clitic "adjuncts" (McCone 1997: 9) and *notae augentes*, illustrated in this subsection with various examples.[11] Independent simple verbs allow a pronominal object, e.g. *-us* in (8), illustrated with the verb *benaid*.[12]

(8) *bentus*
strike$_{3SG.PRES}$-3SG$_{FEM}$
'strikes her'

The alternative strategy is to employ infixation, as in (9), with the pronoun *-m* attaching itself to the available proclitic (here *ní*). Infixed pronouns directly precede the proclitic boundary and come in three classes (GOI §§ 409–427); the choice between class A and B is phonologically conditioned, whereas the choice for class C is conditioned based on syntactic grounds. Simple verbs without a preceding proclitic element acquire the "meaningless" preverbal particle *no* for purposes including infixation of pronouns, illustrated in (10). Infixed pronouns are often accompanied by following initial mutations (which are often not orthographically marked due to the underspecified nature of the Old Irish spelling system).[13]

11 Unless referenced explicitly, examples are either hypothetical or sourced from eDIL.

12 The derivation is *ben(a)ith* + *us* with subsequent syncope (for which see 3.4) and delenition of *th* after *n*.

13 For the initial mutations see GOI (§§ 229–244). Lenition is the pronunciation of consonants with less acoustic energy. As Thurneysen has pointed out, scribal evidence of lenition in Old Irish is initially confined to the letters *p*, *t*, *c* which turn into fricatives, marked by a following *h* (*ph* /f/, *th* /θ/, *ch* /x/). Lenition of *f* and *s* is not indicated in the earlier glosses. Lenited *f* is silent and may be omitted altogether in the spelling; lenited *s* represents /h/. In the course of Old Irish, lenition is also marked on *f* and *s* by employing a *punctum delens* (*ḟ*, *ṡ*). Nasalisation refers to the prefixing of *n* to an initial vowel and the homorganic nasal to *b* and *g* (*mb* /mb/, *ng* /ŋg/), voicing of *p*, *t*, *c* and *f* (hardly ever expressed in the spelling) and gemination with *s*, *r*, *l*, *m*, *n* when preceded by a proclitic vowel (not always marked in the spelling).

(9) *ním·beir*
NEG-1SG·$^{\text{LEN}}$carry$_{\text{3SG.PRES}}$
'does not carry me'

(10) *nom·beir*
PV-1SG·$^{\text{LEN}}$carry$_{\text{3SG.PRES}}$
'carries me'

Special relative endings exist for independent simple verbs for the absolute third person singular and first and third person plural, e.g. (11). In other cases, relativity is marked by an initial mutation following the proclitic preverb, as in (12), or suffixing *-e/-a*, followed by lenition, in case of the preverbs *im(m)* (exemplified in [13]) and *ar*.[14]

(11) *léices*
let$_{\text{3SG.PRES.REL}}$
'who lets, which (s)he lets'

(12) *do·léici*
PV·$^{\text{LEN}}$cast$_{\text{3SG.PRES}}$
'who casts, which (s)he casts'

(13) *imme·thét* (*imbi-tēg-*)
PV-REL· $^{\text{LEN}}$go.about$_{\text{3SG.PRES}}$
'who goes about, which (s)he goes about'

The enclitic *notae augentes* occur in final position in the verbal complex and reinforce an already present subject or object, as in (14).[15]

(14) *at·beir=som* (*ess-ber-*)
PV-3SG$_{\text{NEUT}}$·say$_{\text{3SG.PRES}}$ =3SG$_{\text{MASC/NEUT}}$
'he says it'

We arrive at the schematic overview in (15), loosely based on McCone (1997: 17) (C = conjunct particle, * = zero or more, with the provision that the total of

14 Occasionally *a*/*e* appears with other preverbs (GOI § 493.4): *reme·* (for *remi·*), *íarma* (for *íarmi·*, *íarmu·*) and *assa·* (instead of *as·*).
15 Examples of this form in the glosses are cited in Griffith (2008: 59).

preverbs does not exceed four, E = ending, A = adjunct, N = nota augens; A and N cannot occur together in an independent simple verb).

(15) indep. simple: VROOT E_{ABS} (A) (N)
indep. simple augm.: AUG (A) · VROOT E_{CONJ} (N)
depend. simple: C (A) · (AUG) VROOT E_{CONJ} (N)
indep. compound: PV1 (A) · PV* (AUG) PV* VROOT E_{CONJ} (N)
depend. compound: C (A) · PV1 PV* (AUG) PV* VROOT E_{CONJ} (N)

Taking together all inflectional forms across the tense/mood paradigms, we arrive at about one hundred and twenty inflected forms per verb. If we include affixed adjuncts, augments and *notae augentes*, we are talking about several orders of magnitude more. This "combinatorial" problem is compounded by the fact that scribal practice was often to present the composite elements in the verbal complex as a concatenative string. This results in segmentation challenges, which will be addressed in 7.2.3. Essentially, a computational framework should be able to identify the verb root and all its surrounding elements in strings without mid-high dots, spaces and hyphens, as in *nondobmolorsa* in (16), found in the Würzburg Glosses (Wb.) (*Thes*. 1: 593). This example contains the first singular present indicative of the deponent verb (see 3.3) *molaithir* 'praises' with a *nota augens*, preceded by the meaningless preverbal particle *no*. In the indicative and subjunctive tenses, *no* is used to infix relative *n*,[16] signalling a nasalising relative clause (GOI §§ 497–504). This is what we have in (16), enforced by the conjunction *hore* (*óre, [h]úare*) 'because' and realised by nasalisation of initial *d* of the infixed pronoun.[17]

(16) *hore no -n -dob mol -or -sa*
PV -NAS -2PL · praise -1SG.PRES =1SG
'because I praise you' (Wb. $14^{c}18$)

16 http://dil.ie/33202.
17 For the initial mutation called *nasalisation* see fn. 13.

3.3 A brief description of stem and ending formation

Apart from the rather small class of hiatus verbs, with roots ending in a vowel, Old Irish exhibits an opposition of weak (W1/W2) and strong verbs (S1–S3), which are classified according to present stem formation.[18] Verbs have five stems: present, subjunctive, future, preterite and preterite passive. Stem formation with weak verbs is through largely regular and hence predictable suffixation. Strong verbs show a combination of suffixation, vowel alternations (ablaut) and reduplication, which are largely unpredictable unless one knows the underlying abstract root shape (Stifter 2009: 96). For instance, *crenaid* 'buys' has a future 1sg. conj. *·cíur*, which can be explained by reduplication of the abstract root *cri* → *ci-cr* ... and subsequent lengthening of *i* to compensate for the disappearance of lenited (fricativised) postvocalic *c* before *r* (GOI §§ 71, 691). While Old Irish verb morphology abounds in complex allomorphic stem alternation, further complicated by analogy (for an example see 7.2.2), the term "irregular" is arguably best reserved for suppletion, i.e. usage of different roots across a verb's paradigm.

There are six groups of ending sets which are not arbitrarily combinable with the five stems (Stifter 2009: 96). Apart from the imperative and "secondary" endings (used with the imperfect, past subjunctive and conditional), all ending sets come in two series, i.e. absolute and conjunct, albeit only relevant for simple verbs (see 3.1). Both suffixation and stem-internal modifications are employed in ending formation. The latter comprise alternation of the root-vowel, the change of quality ([non-]palatalisation) of the root-final consonant and the insertion of *u* into the stem ("*u*-infection", Stifter 2009: 67).

There are separate inflectional endings known as deponent, used with a limited set of verbs. While appearing as passives due to endings in *-r*, deponent verbs convey active meaning; the deponent property is therefore "merely lexical", and, consequently, "has to be known for each verb separately" (Stifter 2009: 87).

18 The classification system is the one used in McCone (1997). GOI employs A for weak (and hiatus) verbs, and B for strong verbs (with further subclassifications using Roman numerals). McCone's classification is used here as the letters W and S are more obvious designators for verb type, and a third letter H is reserved for hiatus verbs. Furthermore, McCone's classification reflects a re-examination of inflectional patterns, more clearly showing similarities between inflectional classes (using a subclassification systems of Arabic numerals followed by (optionally) the letters *a*, *b* and *c*). A conversion table is found in Stifter (2006: 381), who also adopted McCone's classification system.

3.4 Syncope

Syncope is the deletion of vowels in even-numbered, non-final syllables in words with more than two syllables (Stifter 2006: 49). In verb forms, the syncope rule operates from the first stressed syllable onwards, that is, the one immediately following the proclitic juncture of the verbal complex. The addition of a *nota augens* (see 3.2) does not cause syncope. The effects of syncope are most pronounced in compound verbs (GOI § 107), where alternation in stress causes much allomorphic variation in the verb stem and the preverbs (see 3.1), e.g. (viii.) *ní·tar†laic*[19] (*to-ro-lēc-*) in Table 2, with deletion of *o* in *ro*. A syncopated front vowel (*e, i*) results in a palatalised consonant cluster, while a syncopated back vowel results in a non-palatal cluster. The latter explains the surface form *·tarlaic*, where the consonant cluster *-rl-* becomes non-palatalised because the syncopated vowel was a back vowel, with verb root *lēc* surfacing as *laic* /ləg^{j}/. There are many attested instances of irregularly applied syncope; an in-depth discussion of some irregular patterns is provided in Ó Crualaoich (1999); see also 7.2.2 in the present chapter.

4 The verb in Middle Irish, and beyond

The Old Irish verbal system undergoes major changes in Middle Irish, eventually resulting in a much-simplified inflectional system in Modern Irish. The key Middle Irish developments are documented in detail in Breatnach (1994a: 278–325) and McCone (1997: 163–241). The changes between Early and Modern Irish are summarised in a–c below.

a. Development of an immutable root shape and transparent stem formation, i.e. univerbation of compound verbs and, as mentioned in 3.1., the gradual development of *ro* as a conjunct particle (Breatnach 1994a; McCone 1997).
b. Replacement of affixed pronominal objects by independent object pronouns (Breatnach 1994a; McCone 1997);
c. Homogenisation (and later renewal) of personal endings, the gradual emergence of independent subject pronouns (outside copula constructions) and, in conjunction with this, analytic verb forms (Breatnach 1994a; Greene 1958; Greene 1973; McCone 1997; McManus 1994).

Developments a. and b. reach completion in Middle Irish, while the developments in c., apart from the streamlining of present and preterite endings, are present in

19 The dagger denotes a syncopated vowel.

embryonic form (subject pronouns) or take place, for the most part, in Early Modern Irish (the development of analytic verb forms). A comprehensive discussion of these pervasive changes is outside the scope of this paper, but some important references have been provided.

The opposition of deuterotonic and prototonic and associated morphophonemic variation was largely done away with by creating new (generally weak) simple verbs based on mainly old prototonic compound bases (McCone 1997: 192–193). This can be illustrated with *do·léici*, prototonic *·teilci*, developing into the simple verb *teilcid* on the basis of analogy with the simplex: *léicid*: *léici*, x: *·teilci*, x = *teilcid*. A more extreme example of stem simplification is the Old Irish compound verb *do·sluindi*, *·díltai* (*dī-slond-*) 'denies', developing into the Middle Irish simple verb *díltaid* (McCone 1997: 207–209), which is the basis for the modern stem *diúltaigh*. These examples illustrate how a verb stem or lemma can change beyond recognition between Old and Modern Irish.

5 Survey of digital resources and computational methods

5.1 Overview

This section gives a survey of resources and tools to be incorporated in – or potentially useful for – the author's research. The main goal is to illustrate the under-resourced status of historical Irish. The introduction to the present volume already documents the available lexical resources and corpora for Early Irish. This section, therefore, focusses on resources for Modern Irish. The most important digital resources are plotted on a timeline in Figure 1. A distinction is made between lexicons and corpora, which are discussed separately in 5.2 and 5.3, respectively. The picture that emerges is one of fragmentation and, especially in the case of lexicons, discontinuity. We are faced with a "lexicographical gap" in the middle, roughly corresponding to the Early Modern Irish period (13th–mid-17th centuries). Discussing modern scholarship and bardic poetry, Mac Cárthaigh (2018: 28) observes that "we still lack such basic infrastructure as a dedicated dictionary for the [Classical Modern Irish] period". Similar observations have been made in Griffith, Stifter, and Toner (2018), who provide a comprehensive research survey on Early Irish lexicography. Subsection 5.4 provides a short excursion into Natural Language Processing for historical texts, and efforts made so far in this area in the Irish context, paving the way for the author's proposed methodology in section 6.

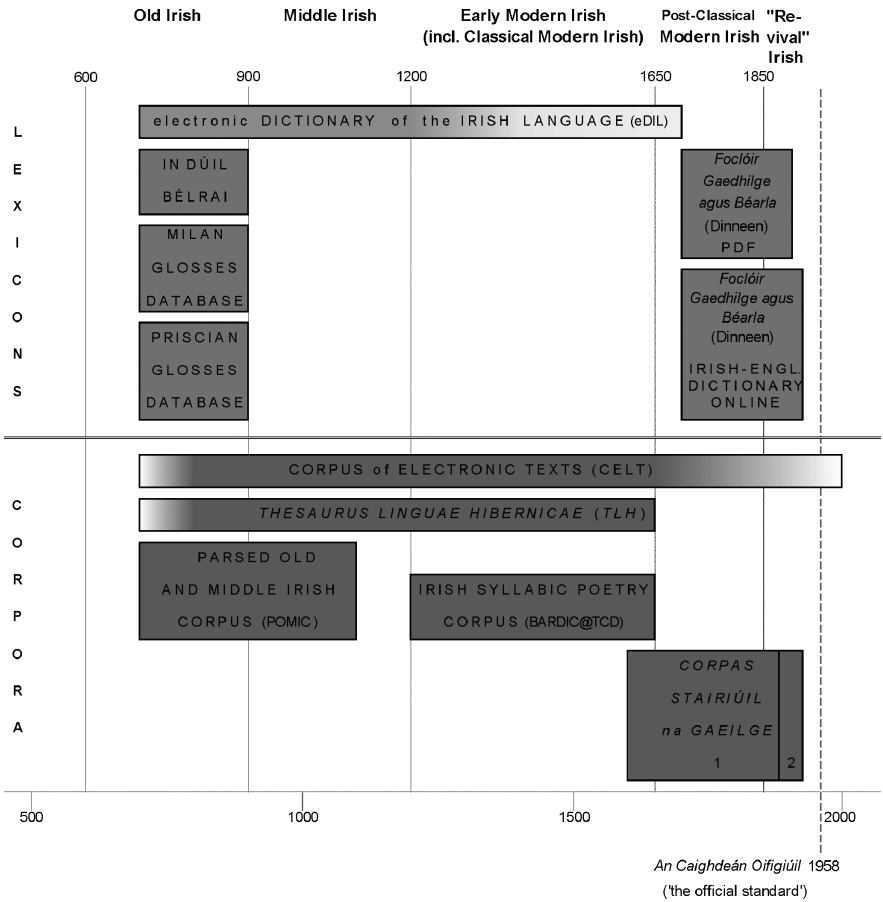

Figure 1: Visual representation of available digital linguistic support for historical periods of Irish. Lighter shades denote lesser support.

5.2 Lexicons

The Dictionary of the Irish Language (DIL) is the only dictionary that bridges the Early and Modern Irish period and "its publication as an electronic resource has been a great boon" (Stifter 2009: 59). However, the resource is not an ideal starting point for an Old Irish morphological parser due to aspects of structure and contents, inherited from the original hard copy. For example, the dictionary is far from exhaustive in listing inflected forms. Other limitations, some of which have meanwhile been remediated by the publication of the electronic

version, are discussed in Nyhan (2006). It should be added 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).

The most important dictionary for Post-Classical Modern Irish is Dinneen (1927), digitised versions of which were prepared in the context of a few independent projects. Publicly available resources include a PDF version of the first edition of the dictionary[20] as well as the online Irish-English dictionary,[21] the latter allowing both English and Irish searches, including the option to be directed to the relevant scanned page of the 1927 edition. The research goals of another project, Digital Dinneen, bear resemblance to the goals of the present work. The aim of this unfinished and dormant project was to create an edition that could be integrated with (mainly) Early Irish resources, including an XML-encoded electronic Lexicon of Medieval Irish (Nyhan 2006),[22] eDIL and CELT. The resulting infrastructure was envisaged to allow a user to follow a Modern Irish form back to its earlier forms (Nyhan 2008). No tools were implemented, but the Digital Dinneen project has produced a (not publicly available) XML-encoded version of Dinneen (1927).[23]

5.3 Corpora

The Irish Syllabic Poetry (or "Bardic Poetry") corpus (c. 1200–c. 1650) consists of approximately 2000 poems from the Classical Modern Irish period, including 500 previously unpublished ones edited in McManus and Ó Raghallaigh (2010). Corpus preparation and annotation is a joint effort by the Irish Department in

20 Available at: https://celt.ucc.ie/Dinneen1sted.html [accessed 30 January 2019].

21 Available at http://glg.csisdmz.ul.ie/index.php [accessed 30 January 2019].

22 While extensively documented in a Ph.D. thesis, this resource is, unfortunately, not available. The following link to a sample of the Lexicon was kindly provided to me by Peter Flynn (email dated 4 November 2014), former manager of the Academic and Collaborative Technologies Group (ACTS), University College Cork IT Services: http://research.ucc.ie/lexicon/sample [accessed 4 May 2019].

23 For online information about this project, see https://celt.ucc.ie/digineen.html [accessed 7 February 2019]. Further information was obtained by means of email contact with Beatrix Färber (30/10/2014 and 03/11/2014), who had the initial idea for *Digital Dinneen*, and Julianne Nyhan (23/2/2012), who informed the present author that neither a lookup mechanism nor a search interface has been implemented.

Trinity College Dublin, the School of Celtic Studies (Dublin Institute for Advanced Studies) and Dr Katharine Simms of the History Department in Trinity College Dublin, who indexed the poems and has created a database, which is currently being updated.[24] As part of the new project BARDIC@TCD (Eoin Mac Cárthaigh and Elaine Uí Dhonnchadha), a POS-tagged corpus currently consisting of 500 syllabic poems has been made freely available and it will be updated regularly.[25]

The tagging of the above-mentioned Bardic Poetry corpus employs automatic standardisation techniques which had already been developed in the context of *Corpas Stairiúil na Gaeilge* 'Historical Irish Corpus' (envisaged to comprise 90+ million words), constituting the basis for the Royal Irish Academy's ongoing project *Foclóir Stairiúil na Gaeilge* 'The Historical Dictionary of Irish' 1600–2000.[26] Uí Dhonnchadha et al. (2014) report on the adaptation of the "modern" tagging tools for the second segment of this corpus (1882–1926), containing seven million words, many of which in a pre-standard orthography (before 1958; see also 5.4).

5.4 Natural Language Processing methods

Natural Language Processing (NLP)[27] is concerned with the ability of computers to process human language (Jurafsky and Martin 2009: 35). The NLP pipeline involves the following activities (in this order): tokenisation,[28] lemmatisation,[29] part-of-speech (POS) tagging[30] and syntactic parsing.[31] A crucial activity in the case of historical texts (and non-standard language in general) is spelling normalisation, influencing all further language processing tasks

24 Available at: https://bardic.celt.dias.ie/ [accessed 7 February 2019].

25 The project website with a link to the corpus is found at https://www.tcd.ie/slscs//research/areas/corpora/bardic.php [accessed 17 July 2020].

26 Available at: https://www.ria.ie/research-projects/focloir-stairiuil-na-gaeilge. The corpus is found at http://corpas.ria.ie/ [accessed 7 February 2019].

27 Alternative names for the field are Speech and Language Processing, Computational Linguistics and Human Language Technology.

28 Separating punctuation marks and other non-alphabetic characters from words (Jurafsky and Martin 2009: 67).

29 Grouping inflected forms of a word under its base form, i.e. its lemma (Mitkov 2003: 744).

30 Assigning a syntactic class marker (e.g. verb, noun) to each word in a corpus (Jurafsky and Martin 2009: 167). POS taggers may be rule-based or trained on annotated data (e.g. statistical), or both.

31 Parsing is a broadly defined concept in Speech and Language Processing that involves taking an input form and produce a structured linguistic representation. Parsing can be done on the morphological, syntactic, semantic and discourse level (Jurafsky and Martin 2009: 79).

(Piotrowski 2012: 11). Standardisation of historical forms to arrive at modern forms is best described as spelling modernisation (Piotrowski 2012: 69–70). The term "canonical cognate" is used by Jurish (2010) to refer to the mapping of an extant equivalent of a historical word that preserves the latter's morphological root and morphosyntactic features. However, sometimes the aim is not to map a historical form to a modern form, but instead to a normalised or canonical historical spelling. This typically involves dealing with both diachronic and synchronic variation.

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

> [T]here is no underlying computational model that describes how synchronic and diachronic variants relate to each other and – possibly – to some shared meaning or some kind of prototype that represents the relatedness of the variants (Piotrowski 2012: 9)

Piotrowski (2012) has pointed out that historical language is inconsistent and highly variable, which hinders POS tagging. The same author mentions various way of tackling this problem. Two common methods, often used in conjunction with each other, are:

a. Bringing an older language variety in line with a standardised or normative – typically modern – variety (either by using rule-based or statistical methods) and use a "modern" POS tagger, if it exists.
b. Employing already existing lexical resources, and create mappings across resources, i.e. through lemmas, dictionary headwords, etc.

In the *Foclóir Stairiúil na Gaeilge* 'The Historical Dictionary of Irish' project (1600–2000), a morphological analyser and POS tagger for the standard language (Uí Dhonnchadha and van Genabith 2006) are conjoined with a standardiser (*An Caighdeánaitheoir* [Scannell 2009, 2017]),[32] employing rule-based and statistical methods and a lexical database of historical and modern word pairs, created by the project's language experts (Uí Dhonnchadha et al. 2014). Initial evaluation of the POS tagging of the 1882–1926 segment of the corpus pointed to F-scores[33] ranging from 91–96% (Uí Dhonnchadha et al. 2014).

Dereza (2018), who discusses lemmatisation approaches for ancient and morphologically complex languages, reports that neither rule-based approaches

32 Code available at: https://github.com/kscanne/caighdean/ [accessed 10 March 2020].

33 A measure of a test's accuracy that incorporates "precision" (e.g. what percentage of the items subjected to standardisation were correctly standardised?) and "recall" (e.g. what percentage of items which should have been standardised were actually standardised?) (Jurafsky and Martin 2009: 479).

(using stems and affixes) nor statistical machine learning methods are useful for Early Irish due to morphophonological complexity, non-transparent orthographical features and scarcity of data. She has developed an Early Irish lemmatiser using *form:lemma* mappings extracted from eDIL and compared two methods: 1) an approximate matching approach using a lemma predictor based on the Damerau-Levenshtein distance, checking for all possible strings of the forms on edit distance 1 and 2,[34] and 2) a neural network approach learning character-level sequences.[35] The first implementation of the lemmatiser shows 45.2% accuracy (i.e. the percentage of correctly generated lemmas) with unknown words and 71.6% with known words, while the neural network metrics are 64.9% and 99.2%, respectively; the neural network approach thus greatly outperforms the one based on edit distance.

6 Proposed methodological framework

In this section, the author will briefly point out how the interlinking of cognate verb forms is envisaged (see section 8), and how some of the resources described in section 5, together with a morphological analyser for Old Irish (section 7), will be employed to this end. The project's methodological framework is schematically represented in Figure 2.

Two morphological finite-state transducers (FSTs, see 7.1), located at the opposite ends of the chronological spectrum, play a pivotal role in the envisaged mapping of cognate historical (verb) forms. Both Old Irish and contemporary standardised Modern Irish reflect stable and normative phases in the language's history and are (relatively) well resourced. For the modern standard language, a morphological FST and a POS tagger are available (Uí Dhonnchadha and van Genabith 2006). As illustrated in Figure 2, standardisation methods are formulated relative to Old Irish and contemporary standardised Modern Irish. Advanced computational standardisation methods are already successfully being used for tagging *Corpas Stairiúil na Gaeilge* 1600–1926 (Uí Dhonnchadha et al. 2014) and the Bardic Poetry corpus, as discussed in 5.3 and 5.4.

34 Minimum edit distance, an approximate matching technique widely used in Natural Language Processing, calculates how similar two strings are by calculating the minimum number of editing operations (insertion, deletion, substitution, transposition) 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: 74).

35 Available at: https://github.com/ancatmara/early-irish-lemmatizer [accessed 13 February 2019].

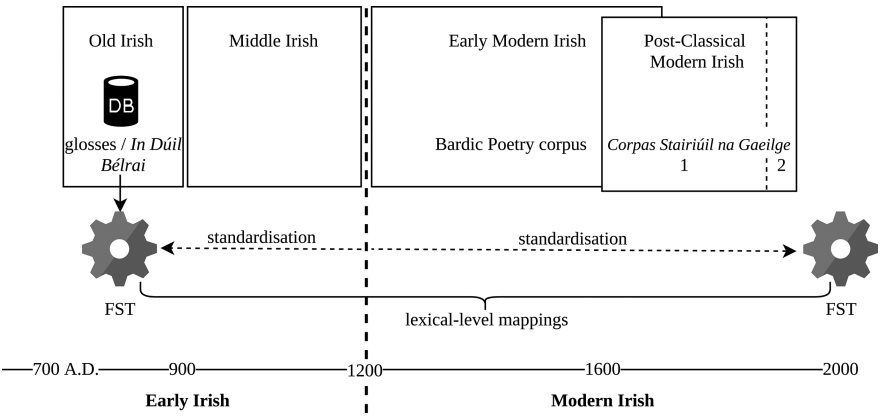

Figure 2: Framework for automatic identification and linking of cognate Irish (verb) forms. FST = finite-state transducer, see 7.1.

Seeing that much progress has already been made in tagging increasingly earlier historical Modern Irish forms, the present author is concentrating on the Early Irish side of the timeline. The following three tasks constitute the framework of the author's project:

a. Building a morphological finite-state transducer (FST) for Old Irish, which can assist in future work on a POS tagger for this period.
b. Creating lexical-level mappings between the Old Irish morphological analyser and the available tagging tools for Modern Irish.
c. Employing standardisation methods and potential analyser/tagger adaptation, in conjunction with digital corpora, to cover the language periods between Old and Modern Irish.

Task a. reflects the most novel approach in the author's research project. The finite-state transducer can be augmented with manually parsed data from the databases of the Old Irish glosses (currently being streamlined in CorPH, see the introduction to this volume) and partial lemmatisation tables for verbs as present in *In Dúil Bélrai* (King, Lash, and Gabay 2006). It should be noted that the present work deals with morphological parsing rather than POS tagging. The task of automatic morphological analysis is to present all the grammatical possibilities on the word level. POS tagging is a subsequent task that aims at disambiguating between morphological parses (e.g. is Old Irish *ben* a verb or noun?) based on commonly the phrasal context. Due to the highly synthetic nature of the Old Irish verb, fine-grained morphological analysis is an essential prerequisite for POS

tagging as well as other subsequent tasks in the NLP pipeline for Old Irish. Morphological parsing of the Old Irish verb is the topic of the next section.

7 Automatic morphological analysis and generation of Old Irish verb forms

7.1 Finite-state morphology

Finite-state morphology is based on the mathematical notion of a finite-state automaton, a machine that recognises a particular set of symbol sequences (strings) as defined by a regular expression (a language for specifying text search strings, Jurafsky and Martin 2009: 17–18). Automata can be conceptualised as networks with transitions through a finite number of states. A finite-state transducer (FST) is an extension of this concept and contains two-level symbol correspondences for each path in the network. Figure 3 shows an FST with a mapping between a lexical-level and surface-level string representing present indicative third person singular absolute *léicid*. One of the advantageous features of this two-level formalism is that the relations encoded are inherently bidirectional: an FST can be used in recognition mode to analyse (parse) orthographical words in a text, but it may also be used to generate, say, a specified set of inflected forms (listing, for instance, complete paradigms of verbs with root *lēc*). Jurafsky and Martin (2009: 80) describe an FST as a "key algorithm for morphological parsing . . . and crucial technology throughout speech and language processing".

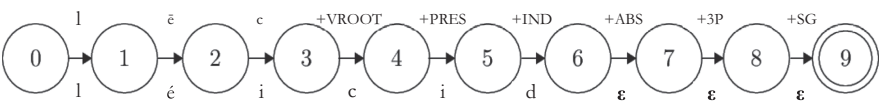

Figure 3: A finite-state transducer accepting, at final state 9, the surface string léicid (lower level) and lexical string lēc+VROOT+PRES+IND+ABS+3P+SG (upper level), constituting a two-level mapping. The epsilon (ε) denotes a so-called "empty transition": a mapping where there is no accompanying symbol on the opposite level, i.e. when the upper and lower strings are of unequal length. The term analysis is used for upward mapping, which translates into morphological parsing. Downward mapping equals generation of (most commonly) orthographical strings.

Beesley and Karttunen (2003) is an important reference work accompanied by a toolkit called Xerox Finite State Tool (*xfst*).[36] This tool provides an extended set of regular expression operators, including the conditional rewrite rule format used in phonology, to intuitively model morphological and morphophonemic processes. The lexicon compiler (*lexc*) program (Beesley and Karttunen 2003: 203–278) facilitates and simplifies the creation of morphological grammars and can be used in conjunction with *xfst*. The finite-state toolkit *foma* (Hulden 2009) – which is freely available[37] and compatible with *xfst* – is used by the author to develop a morphological FST for Old Irish. The development of this tool is extensively documented in the author's Ph.D. thesis (Fransen 2019), which, together with the code, is available online.[38]

7.2 Implementation: Some highlights and challenges

7.2.1 Lexical and surface-level description

The two-level morphology paradigm is a fitting choice for the often daunting discrepancy between underlying and surface forms in Old Irish (verb) morphology, as detailed in section 3. The observation below is a suitable precursor to the computational challenges faced and choices made, as detailed in the remainder of this section:

> The bewildering complexities . . . become transparent only when viewed from a diachronic position, and in order to understand allomorphic variation correctly it is essential to work with underlying forms and their often quite dissimilar surface representations
>
> (Stifter 2009: 60)

The two-level formalism does not prescribe which linguistic entities are to be assigned to the upper level, although the latter is commonly reserved for synchronically motivated underlying morphemes. The (final) surface-level forms, however, should obviously match against the (commonly) orthographical forms as found in a text corpus.

Typically, the lexical level starts off with a lemma, and the surface level with a stem. In many languages, the latter bears an obvious relation to the former. This relation, however, is far from trivial in Old Irish, as pointed out by

36 The accompanying website is http://www.fsmbook.com [accessed 7 February 2019].
37 Available at: https://fomafst.github.io/ [accessed 25 January 2019].
38 Available at: https://github.com/ThFransen84/OIfst [accessed 10 March 2020].

Stifter (2009: 60). The full range of surface (inflected) forms often cannot be deduced from a single inflected form across a verb's paradigm, which means that the conventional citation form provided in lexicographical resources and grammatical descriptions of the language, the independent present indicative third person singular (e.g. eDIL),[39] is of little use when it comes to formulating a surface-level stem entry. This hurdle will be tackled in the next subsection.

The lexical level in the author's FST consists of diachronically motivated underlying morphemes rather than a citation form. In other words, a verb form's upper-level parse includes the abstract root shape of the verb and (with compound verbs) the underlying form of the preverb(s). In addition to linguistic motivations echoing Stifter (2009: 60), there are two practical reasons for the author's *modus operandi*. First, the use of diachronically motivated verb roots enables one to generate (surface) verb forms which have the same historical verb root. Secondly, employing "diachronic tags" allows for – and facilitates – interoperability with projects dealing with other historical Indo-European languages, or, indeed, Proto-Indo-European.[40]

Example (20) illustrates the two-level encoding of the verb form *as·oilgi* 'opens' (L = lexical level, S = Surface level), based on the derivation provided in Stifter (2006: 364). The lexical-level tag *+PROCL_JUNCT* denotes the separation between the proclitic(s) and the stressed part of the verbal complex. The upper-level tag W2a can be added to enable extraction of verbs with this specific stem type. Consecutively numbering the preverbs is also expected to facilitate in-depth linguistic analysis; for example, it allows for a systematic investigation of the positional hierarchy of preverbs.

(17) L uss+PV1+PROCL_JUNCT+od+PV2+lēc+VROOT+W2a+PRES+IND+CONJ+3P+SG
S *as·oilgi*

Even though Old Irish can be treated as a normative phase within the medieval period, the language is far from orthographically stable. In the current

39 It should also be noted that headwords in eDIL are not consistently provided in a form representative of Old Irish. An example is classical Old Irish deponent *molaithir* 'praises', which is represented by "generic" Early Irish *molaid* in the dictionary. (a new entry, albeit solely containing a reference to molaid, has been introduced in the revised 2019 version of the dictionary; *s.v. molaithir* or dil.ie/50393).

40 See, for example, *Proto-Indo-European Lexicon*, a generative etymological dictionary of Indo-European languages, also implemented with the finite-state toolkit *foma*. Available at: http://pielexicon.hum.helsinki.fi/[accessed 7 February 2019].

implementation, the surface or lower level adheres, as closely as possible, to classical Old Irish grammar and orthography. Orthographical variation in Early Irish texts is expected to be successfully handled by one of Dereza's (2018) lemmatiser implementations (see 5.4) used as a standardiser (see 7.4).

7.2.2 Monolithic stems

Section 3 has illustrated that a significant amount of allomorphic variation can be seen with verb stem formation, with syncope often causing truncation of the verb root, as in (17) above. This variation is challenging for a finite-state rule-based system, in which one typically starts off with a list of stems and affixation rules. Recall the morphotactics of the verbal complex (see example 7 in 3.1), repeated here as (18).

(18) (C) PV* (AUG) PV* VROOT E

If we blindly applied the morphological concatenations without regard to phonology, we would get, for instance, *ní-to-ro-lēc-* (C-PV1-AUG-VROOT), where the morphological derivation is quite far removed from the surface or orthographical form *ní·tarlaic* (see [viii.] in Table 2 on page 55, and 3.4). Employing the above concatenation schema to model Old Irish verb morphology was therefore not considered a feasible starting point – even when equipped with knowledge about the positional hierarchy of preverbs (McCone 1997: 89–90).

Allomorphic alternations are essentially a product of the morphology-phonology interface in Old Irish, as has been demonstrated in section 3. In other words, "unpredictable" stem formation is largely due to stress patterns, including syncope. Looking at Table 2 in section 3.1, the examples that do not show allomorphic stem variation are simple verb forms (i.)–(iii.) and the deuterotonic compound with one preverb in (v.), exactly those forms that have a stressed verb root. In all the other examples given in Table 2, where the root is unstressed ([iv.] and [vi.]–[viii.]), stem formation is less trivial, at least from a computational viewpoint and if operating with a set of clear-cut, synchronic rules.

The opposition of stressed versus unstressed verb root was found to be of major significance in formulating verb stems.[41] An additional base form is required for any combination of a preverb or augment and an unstressed verb

41 Note that this distinction does not fully coincide with the traditional binary oppositions of simple versus compound and independent versus dependent.

root. A simple weak verb often requires an additional entry for the dependent augmented form, such as *reilic* in the case of *léicid*. A weak compound verb requires more stems; *do·léici*, for example, can be said to have four stems: (*do*) *léic*, (*do*)*reilic*, *teilic* and *tarlaic* (see Table 2). While stems of weak verbs are generally unmodified in the different tenses/moods, strong verbs may show root-internal stem modifications in each of the five tense/mood stems; in other words, the above-mentioned numbers for weak verbs should be multiplied by a factor of up to five for strong verbs.[42]

Stem entries such as *reilic*, *teilic* and *tarlaic* are called "monolithic stems" in the author's computational framework. These bases represent synchronically motivated multi-morpheme strings not trivially segmentable on the surface. Accordingly, they are not produced by diachronic phonological rules in the author's FST rule framework, but keyed in as invariant stems in the *lexc* grammar. Monolithic stems subsequently enable the encoding of straightforward inflectional endings.[43] While initially born out of programming considerations, the concept of a monolithic stem is perhaps also theoretically insightful. When these bases have been determined and encoded for a large amount of verb lemmas, the minimum or average amount of stems necessary for operating with simple morphological rules can be calculated, which could be an interesting linguistic diagnostic for the level of complexity of the Old Irish verbal system.[44]

The formulation of monolithic stems partly remedies the problem of synchronically opaque stem formation and alternation. However, dealing with syncope remains a complicated aspect in the implementation. For example, syncope may cause secondary palatalisation/non-palatalisation. Consider the independent (deuterotonic) and dependent (prototonic) present indicative first person plural forms

42 The concatenation of preverbal proclitics results in less allomorphic variation and can be modelled using separately defined surface morphemes (more on the programmatic treatment of proclitics versus (stressed) stems in 7.2.3).

43 But see the discussion on complications associated with syncope below.

44 I am thankful to Prof. David Stifter for bringing this additional insight to my attention. One could ask the question how the complexity in Old Irish verb stem formation compares to other languages. Such a cross-linguistic examination is, unfortunately, outside the scope of this paper. It is not unlikely, however, that the perceived complexity of the verbal system is at least in part due to the absence of a comprehensive synchronic description of Old Irish, and, consequently, a framework employing transparent morphological rules. So far, scholars of Old Irish have been mainly relying on historically oriented grammars such as GOI. A related issue, not unique to Old Irish, is the fact that the description of a historical language is typically based on a closed and often relatively small corpus; many forms across inflectional paradigms are not attested, which may impede a full synchronic description of the morphological rules at play.

of the strong verb *as·beir* (*ess-ber-*) 'says' in (19) and (20), respectively. The root vowel *e* in *ber* has been subject to syncope in (20). The *e* in ending *-em* in (20), as opposed to *-am* in (19),[45] marks subsequent secondary palatalisation of the consonant cluster *-pr-* (/bʲrʲ/). Example (21) is the dependent prototonic equivalent of *as·beir*, with palatalisation of root-final *r* (/ebʲərʲ/) to mark the personal ending. This form is not liable to syncope as it only consists of two syllables. The mechanisms behind stem and ending variation of this kind occur throughout paradigms of compound (and augmented simple) verb formations.

(19) *as·beram*
PV·say$_{1PL.PRES}$
'we say'

(20) *ní·eprem*
NEG·say$_{1PL.PRES}$
'we do not say'

(21) *ní·epir*
NEG·say$_{3SG.PRES}$
'does not say'

In the current implementation, syncope is incorporated in the framework of regular expression rules; a conditional rewrite rule targets vowels in even-numbered syllables (but not final ones), which are liable to syncope. A monolithic stem such as *tarlaic* should therefore be encoded as *tarolaic* (even though this form never surfaces) to make sure vowel syncope is correctly applied in subsequent even-numbered syllables. Monolithic stems are therefore perhaps best described as semi-surface forms.

Unavoidably, "mechanical" treatment of syncope results in cases where the resulting consonant cluster violates the phonotactics of Old Irish. While this can be (and partly has been) counteracted by changing the conditional rewrite rule, irregularly applied syncope is very hard to cater for. For example, the augmented preterite third person plural surface form *reilciset* (*·reil†ciset*) generated by the FST does not match attested *·rel†c†set*,[46] with syncope of the vowel in the second as well as third syllable (the difference between the vowel sequence

45 Apart from the different quality of the preceding consonant(s), both endings represent /əṽ/.
46 This form is cited and discussed, alongside other examples of compounds with root *lēc*, by Ó Crualaoich (1999: 97–98).

e and *ei* in the first syllable is purely orthographical). Perhaps we should operate with the stem *reilc* instead of *reilic* to arrive at what would then be regular syncope of the vowel in the second syllable; however, the question in that case is how to derive forms without syncope, such as expected and attested[47] preterite third person singular *·reilic*. Intra-paradigmatic analogy further complicates a rule-based approach to syncope, as can be illustrated with dependent present third person plural passive *·epertar* (expected *·*ep*†*retar*) of *as·beir* 'says' (*ess-ber-*), modelled on the present third person singular passive *·eperr*.[48]

The complexities relative to syncope and analogy (operating both within and across paradigms) raise the question whether rule-based stem-and-ending generation using monolithic stems is invariably more economical than manually encoding ("hard-coding") an entire verb paradigm. Strong verbs such as *beirid* and *as·beir* are very frequent, and therefore more liable to irregularity and analogical processes. For many other verbs, the distinction between "regular" and "irregular" (or, perhaps better, "predictable" versus "unpredictable") is not as clear-cut, which deters deciding *a priori* whether an automatic generation or manual encoding approach is most feasible. Establishing a good balance between automatic and manual methods (based on expert knowledge) is further complicated by the fact that no exhaustive list of Old Irish verbs or verb roots exists – let alone a comprehensive overview of stem classification and stem formation processes that could inform the formulation of monolithic stems.[49]

In the author's project, the focus is initially on weak verbs; compared to the group of strong verbs, weak verbs show transparent tense/mood stem formation

47 eDIL *s.v. léicid* or dil.ie/29766.

48 For an overview of the entire inflectional paradigm of *as·beir* see Strachan (1929: 68–71).

49 Another problem is that some works deal with roots, and others with lemmas. Pedersen (1909–1913, vol. 2) lists 204 roots (based on the dedicated number of paragraphs, 650–854). However, the focus is on primary verbs (mainly verbs with Proto-Indo-European roots), which are mainly strong verbs. A more up-to-date work on primary verbs is Schumacher (2004), who lists 197 reconstructed Celtic verb roots, 166 of which are found in Irish verbs. However, this work excludes causatives. Le Mair (2011) discusses weak verbs in the Old Irish glosses, giving a total number of 365. McCone (1997) lists a good number of inflections in his *index verborum* but does not include stem class and mainly considered material from the Old Irish glosses. The online eDIL contains 4,127 verb headwords but does not systematically provide a stem classification. This number includes duplicates as some verbs have a separate Old and Middle Irish headword. Moreover, some (e)DIL headwords are more indicative of Middle Irish than Old Irish (e.g. *molaid*, rather than Old Irish deponent *molaithir* 'praises'). Rossiter (2004) applied McCone's stem classification (the one adhered to in the present work) to verbs in DIL, but only dealt with compound verbs. The vocabulary section in Stifter (2006) is far from exhaustive but does systematically provide the stem class and roots for verbs.

by means of suffixation only (see 3.3); in other words, one does not have to cater for non-trivial stem-internal (non-concatenative) modifications based on an abstract root. However, as the above has shown, most verbs need more than one monolithic stem regardless if we want to cater for augmented simple verbs and compound verbs.

7.2.3 "Word" segmentation and separated dependencies

Morphological parsing operates on the word level, and words are defined as strings surrounded by space. A form such as *beirid*, with the ambiguous ending *-id*, will receive three grammatical analyses during morphological parsing, as it may occur in three grammatical contexts, illustrated in (22)–(24). The presence or absence of a conjunct particle (here negative or negative imperative), if separated by space, is a disambiguating feature in the subsequent task of building a POS tagger (not part of the present paper), which operates beyond the word level, even if merely typographical.

(22) *beirid*
carry$_{3SG.PRES}$
'carries'

(23) *ní beirid*
NEG·carry$_{2PL.PRES}$
'you do not carry'

(24) *(ná) beirid (. . .)!*
(NEG·) carry$_{2PL.IMPV}$
'(do not) carry (. . .)!'

Morphological boundary markers, including spaces, are absent in faithfully transcribed texts and diplomatic editions. More commonly, text editions (which might make their way into a digital corpus) are subject to editorial choices and policy, according to which typographical morpheme boundary markers might be employed. The current finite-state implementation anticipates instances of a potentially highly synthetic verbal complex written as one "word" (consecutive string), optionally with a mid-high dot (for the proclitic juncture) or hyphens; it accepts forms of the type *nondobmolorsa*, but also, for instance, *nondob·molorsa* (example (16) discussed in section 3.2) (a different yet interesting approach focused on pre-processing is provided in Doyle, McCrae, and Downey (2019),

who explore the possibilities of automatic tokenisation for Old Irish using a neural-network-based approach.). This choice in the implementation facilitates recognition but obviously also results in a vast amount of combinations to be considered of which only a limited amount are morphotactically valid. The restrictions are generally separated dependencies (co-occurrence of non-consecutive morphemes) and most of these have been successfully encoded. The generation of exclusively morphotactically valid forms prevents wrong parses due to ambiguity at the surface level (e.g. identical absolute and conjunct endings).

The interaction between monolithic stems (see 7.2.2) and separated dependencies is illustrated in Table 3, using the same verbs as in Table 2.[50] By their very nature, simplexes such as *léicid* cannot be preceded by a lexical preverb. Compound stems of the type *teilic* or *tarlaic* preceded by the proclitic augment *ro* or the proclitic preverb *do* are impossible as one or both of these elements are already incorporated in the (monolithic) verb stem.[51] However, *ní* (and conjunct particles in general) can precede any stem except deuterotonic stems.

The author's *lexc* implementation contains separate lexicons for proclitics (preverbs and particles, optionally with infixed pronoun and relative marker) and verb stems with endings, which may occur typographically as strings separated by space, and are recognised as such. The lexicons can be optionally concatenated. In the author's implementation, separated dependencies are partly encoded with "flag diacritics" (Beesley and Karttunen 2003: 339–373), special regular expression symbols accompanying morphemes (*lexc* entries) that either allow or block paths in the network. Flag diacritics are not visible during analysis/generation, apply at run-time, and can, together with the blocked morphotactically illegal strings, be deleted from the network. For example, if *do* is marked as "preverb *do* seen", and prototonic *teilc* as "preverb disallowed", we will never get, for instance, **do·teilic*. A simple verb such as *marbaid* 'kills', also accompanied by "preverb disallowed", will equally never be prefixed with *do* (or any other preverb). Deuterotonic stems,

50 It should be noted that there might be overlap in monolithic stems across verb lemmas (e.g. *reilic*). In the current implementation, formulation of monolithic stems is on a per-verb (lemma) basis. An approach whereby monolithic stems are used for multiple lemmas, while not impossible, fails to make a distinction between, for instance, simple and compound verbs, which are subject to different constraints. "Recycling" monolithic compound stems might be of use, however, with verbs liable to preverb alternation (e.g. *in·fét*, *ad·fét* 'relates'), secondary composition (Stifter 2006: 254) and the employment of "dummy" preverbs in Middle Irish (McCone 1997: 194–197).

51 *ro·teilic* and *níro·teilic* for *do·reilic* and *ní·tarlaic*, respectively, reflect a Middle Irish development whereby *ro* gradually assumes the status of conjunct particle (see 3.1 and Breatnach 1994a: 279). The prefixation or infixation of proclitic *ro* with prototonic stems is blocked in the current version of the FST; systematic encoding of Middle Irish features (such as the relaxation of grammatical rules relating to proclitic *ro*) is envisaged as a subsequent adaptation stage of the FST.

Table 3: Schematic overview of separated dependencies with a selection of proclitics and monolithic stems in Old Irish, exemplified with *léicid* and *do·léici*.

Lemma	LEXICON 1				LEXICON 2		
	Proclitic ("prefix")				Monolithic stem		Ending
	Conj. part.	Pre-verb	Aug-ment	Optional infix / rel. marker and/or pronoun class	Simple/ proto-tonic	Deutero-tonic	
léicid (simplex)			*ro*	A or rel.(+C)	*léic-*		abs.
							conj.
	ní			A	*reilic-*		
					teilic-		
do·léici (com-pound)					*tarolaic-*		
		do		A or rel.(+C)		*léic-*	
						reilic-	

however, have a flag diacritic of the type "requires preverb *X*", which only allows the specified preverb, and nothing else, not even a proclitic that is not a preverb. Prototonic/simple stems, on the other hand, allow anything *but* a proclitic preverb, and are (mostly correctly) preceded by proclitics that are not preverbs (e.g. *ní*).

Flag diacritics have proven to be convenient for the encoding of (essentially arbitrary) combinations of proclitic preverb and deuterotonic stem. The disadvantage of flag diacritics, from a programmatic point of view, is the fact that one needs to think carefully about separated dependencies *in advance* when laying down the morphological concatenation architecture. Consequently, "adding Flag Diacritics *post hoc* to an existing system can require non-trivial re-editing of your source files" (Beesley and Karttunen 2003: 340). A sometimes more convenient way of restricting the generation of ill-formed words is the use of upper-level filters (Beesley and Karttunen 2003: 247–255), i.e. specifying incompatible upper-level tags for an initially over-generating *lexc* grammar, and filtering all the illegally formed strings out of the network.

7.3 Preliminary test results

The morphological FST was tested on the Old Irish text *Táin Bó Fraích* [Cattle-raid of Fróech], using the digital version available on CELT,[52] taken from the edition by Meid (1974). The FST was augmented by personal names occurring in the story, a limited set of function words, and the extremely frequent defective verb *ol* 'said'. It turned out that 9.6% of word types (unique forms) were morphologically parsed, with an average, comparable score of 10% for four other Early Irish narrative texts edited by Greene (1955).[53] While the consistency of these scores is a promising result, the main goal of this exercise was to see how the FST would cope with weak verb inflection, which was concentrated on during implementation. It should be noted that weak verbs were found to be rather infrequent; in terms of tokens, W1 and W2a verbs constitute only 8.3% of the total amount of verb forms (excluding verbal nouns) in *Táin Bó Fraích*.[54]

Out of the 50 W1 and W2a inflected forms (types) in *Táin Bó Fraích*, 36 (72%) were found to be correctly parsed. Most of the 14 non-recognised forms either deviate from a "canonical" spelling or show idiosyncratic features that are difficult to capture in general rules. Two verb forms in *Táin Bó Fraích* show grammatical variation that perhaps legitimises a rule; present subjunctive third person singular *forruma* (*fo·ruimi, fo·rumai* 'puts')[55] and preterite passive third person singular relative *arrálad*[56] (*ar·áili* 'arranges') show fluctuation in stem-final consonant quality, which is a feature of W2 verbs (McCone 1997: 27–28).

Spelling variation in and across texts such as the ones considered here are often seemingly trivial, as can be illustrated with verbs with stem *léic-*, which, first of all, may equally be spelled *léc-*. Another feature with the no apparent grammatical implication is the occurrence of the digraph *ll* in all three instances

52 Available at: https://celt.ucc.ie//published/G301006/ [accessed 7 February 2019].

53 Also available on CELT. The individual stories are *Fingal Rónáin* (available at: https://celt.ucc.ie//published/G302011/), *Orgain Denna Ríg* (available at: https://celt.ucc.ie//published/G302012/), *Esnada Tige Buchet* (available at: https://celt.ucc.ie//published/G302013/) and *Orgguin trí mac Diarmata Mic Cerbaill* (available at: https://celt.ucc.ie//published/G100037/) [all accessed 7 February 2019].

54 Middle Irish simple weak formations from original compound verbs were excluded from this count, as the FST does not (yet) deal with Middle Irish, e.g. *fácbaid* and *oslaigid*, from Old Irish *fo·ácaib* 'leaves' and *as·oilgi* 'opens', respectively.

55 The headword *fo·ruimi* is given in Meid (1974) and eDIL (s.v. *fo-ruimi*, *-fuirmi* or dil.ie/24043, under which attested third person singular present indicative *forrumai* is listed).

56 The nasalising relative marker appears on the final consonant of the preverb (*-rr*), rather than on *ál-* (*nál-*), as expected. We would also expect the preverb to appear as *ara-*. Variation of this type would have prohibited successful analysis by the FST in the first place.

of sentence-initial present indicative third person singular *do·léici* in *Táin Bó Fraích*, for what is more commonly a single *l* in this context. Frequent alternations of this kind beg the question if – and to what degree – spelling variation should be encoded as a final module in the FST framework. This and related possibilities will be briefly discussed in the next subsection.

7.4 Standardisation for Early Irish

Beesley and Karttunen (2003: 287–293) recommend building different transducers for different tasks so as to make the parsing pipeline as modular and flexible as possible. Such a pipeline typically includes a "standard" FST, where the surface level represents normative grammatical or orthographical forms. A separate transducer could be devised which can be secondarily invoked when a form does not conform to a standard grammar or spelling. The latter may take care of variation in spelling, as with, for instance, *lécid* for "standard" *léicid*.

The lemmatisation tools being developed by Dereza (2018) are also expected to be of benefit for standardisation purposes. As stated in 5.4, Dereza (2018) has implemented an Early Irish Lemmatiser using two approaches: one method is based on approximate matching using string similarity, the other uses neural machine learning. The first implementation predicts a mapping between an unknown inflected form in a text to a known variant, based on a dictionary of *form: lemma* mappings originally retrieved from eDIL. The second, more start-of-the-art and better performing version employs the latter mappings to learn character sequences in order to produce lemmas it has never seen. The morphological FST for Old Irish, currently being developed by the author, will, in time, surpass the amount of inflected verb forms listed in the Dereza's Lemmatiser dictionary. Moreover, the inflected forms generated by the FST adhere to a large extent to classical Old Irish inflection and spelling. By adding these canonical or "standard" Old Irish forms to the known mappings in the Lemmatiser's dictionary, we not only increase the power of the Lemmatiser enormously, but we can also use this resource as a spelling standardiser, namely, by mapping an unknown variant in a text to a "standard" form from the FST, and retrieve the morphological parse of the latter. Lemmatisation and standardisation methods – in conjunction with the author's morphological FST – have only been tested to a very limited extent.

The author has taken the liberty to use the term "standardisation" in his framework (see section 6), to show the similarity with the approach taken in the *Foclóir Stairiúil na Gaeilge* project. The terms "canonicalisation" or "normalisation"

are perhaps more fitting terms as an absolute standard did not exist in Early Irish, at least not in orthographical terms, not even in the otherwise reasonably homogenous language of Old Irish.

8 Suggestions for linking cognate verb forms

The author's most fundamental envisaged approach is to create operability between his own Old Irish morphological FST and the one for Modern Irish (Uí Dhonnchadha and van Genabith 2006). Such an infrastructure could incorporate mappings between lexical-level tags, i.e. between Old Irish preverbs and verb roots and modern verb lemmas of the type *lēc+VROOT*:*lig+Verb+VTI* and *to+PV1+lēc+VROOT*:*teilg+Verb+VTI*. Additional tag mappings between inflectional categories could be devised, with the provision that there is often a discrepancy between Old and Modern Irish. For example, there is no straightforward modern grammatical category that matches the Old Irish augment. Although the modern past tense in many cases etymologically derives from a perfect construction with the preverbal particle *do* (for earlier *ro*, the resultative augment), it does not inherit either "perfectivity" or "perfect" as a grammatical feature.

Nonetheless, tag mappings of this kind facilitate juxtaposition of Old and Modern Irish paradigms, facilitating research into historical roots and grammatical developments such as innovatory processes in stem and ending formation. The historical connection between lemmas such as *lig* 'let' and *teilg* 'cast' is not present in the modern-language morphological FST. However, this connection can be established by means of *lēc+VROOT*, the "common denominator" for all verbs with this root in Old Irish, for which individual paradigms can be generated. The "modern" analysis additionally tells us that both *lig* and *teilg* can be used transitively (*+**VTI***), a feature which is expected, in many cases, to transfer back to Old Irish.

Another "linking route" is through lemmatisation using *droichead* (Scannell 2018), a digitised version of the mappings between standardised contemporary Modern Irish lemmas from *Foclóir Gaeilge-Béarla* 'Irish-English dictionary' (Ó Dónaill 1977) and eDIL headwords, originally prepared by de Bhaldraithe (1981). Scannell (2018) added POS tags and used the imperative second person singular (as in Ó Dónaill 1977) as the modern lemma rather than the third person present indicative (matching the eDIL headword) in the original list. Since the FST for Modern Irish (Uí Dhonnchadha and van Genabith 2006) employs the lemmas in Ó Dónaill (1977) on the lexical/upper level, and *droichead* provides the corresponding eDIL headword, mappings between any modern standard inflected

form (as well as many pre-standard forms, see 5.4) and the Early Irish eDIL headword can be facilitated. Mappings between eDIL and (increasingly earlier) Modern Irish inflected forms or headwords would be of great benefit to scholars working on texts produced at various stages during the medieval period, who are currently confronted with a vast range of grammatical and orthographical variants while operating with limited lexicographical resources, especially for Early Modern Irish.

9 Synthesis and future work

The aim of the work is to link up lexical resources for the Early and the Modern Irish period. This chapter has identified a lack of digital linguistic resources for the historical Irish period, with a fragmentation and discontinuity in terms of lexicographical support, which makes the aim of the research, the interlinking of cognate verb forms, a far from straightforward process. This challenge is compounded by significant linguistic developments between Old and Modern Irish, mainly in the verbal system, and especially in Middle Irish.

The computational methodology proposed employs two finite-state transducers (FSTs) at the opposite end of the historical spectrum – Old Irish and contemporary standardised Irish – as these language stages represent normative and/or standardised varieties and are well resourced. Advanced methods are being employed in the context of the Royal Irish Academy's *Foclóir Stairiúil na Gaeilge*, using a standardiser in conjunction with a modern-language POS tagger (based on an FST), greatly increasing recognition of increasingly earlier historical variants and connecting the latter with the modern lemma.

The current focus in the author's project is on creating a morphological FST (and, subsequently, POS tagging tools) for Old Irish using the software *foma* (Hulden 2009). The FST is planned to be used in conjunction with a lemmatiser for Early Irish based on eDIL (Dereza 2018), which could be employed to predict canonical Old Irish inflected forms generated by the Old Irish morphological FST for orthographical variants in Early Irish texts, as such functioning as a standardiser.

The challenges relating to a rule-based FST include morphophonemically complex verb stem formation. Allomorphic stem variation and truncation of the verb root, especially prevalent with compound verbs, have been tackled computationally by devising multi-morpheme, non-derived units called "monolithic stems" in the author's work; these bases consist of the verb root and, if present, preverbs and augment following the proclitic juncture. While the formulation of

monolithic stems is time- and knowledge-intensive, the resulting stem entries reduce complex, non-concatenative stem and ending formation to relatively straightforward morphological rules, which can be largely automated. Separated dependencies have been successfully handled with instruments of the finite-state paradigm.

Unpredictable inflectional patterns resulting from irregular syncope and analogy in inflectional patterns challenge a linguistically motivated, rule-based approach. A further issue is the absence of an exhaustive list of Old Irish verbs and information about stem type and stem formation. These conditions make it difficult to exactly establish the balance between automatic methods and manual efforts and expert knowledge needed. However, the concept of a monolithic stem strikes an interesting balance between "automatic" and "manual" and may well be a leap forward in establishing this balance. The incorporation of lexical resources such as the databases produced as part of the *Chronologicon Hibernicum* project and the lemmatised verb tables as part of *In Dúil Bélrai* will likely speed up the development of the author's FST.

Test results are promising but incorporation of more verbs and verb classes as well as catering for inflectional variation and non-standard forms is an important prerequisite in the context of establishing the feasibility of the implementational choices and further development of the FST for Old Irish.

Linking cognate verb forms across the entire historical period is very much future work. However, two methods have been proposed in this work. The first one involves mappings on the lexical level of the FSTs for Old and Modern Irish, facilitating the juxtaposition of entire historical paradigms based on Old Irish roots as well as systematic investigation of linguistic change. Alternatively, mappings between eDIL headwords and modern lemmas from Ó Dónaill (1977) can be established by integrating the tagger (and standardisation tools, Scannell 2009, 2017) for Modern Irish (Uí Dhonnchadha and van Genabith 2006) and the mappings as part of *droichead* (Scannell 2018).

Standardisation methods in conjunction with the Old Irish and Modern Irish morphological analysis/tagging tools will result in increasingly better coverage rates of intermediate variants. With the modern-language tagger "stretching back" and the one for Old Irish "reaching forward" we can metaphorically describe the adaptation process as a "two-pronged attack". It should be stressed that, in catering for variation throughout the medieval period, adaptation processes are likely to move beyond the realm of orthography.

The substantial linguistic variation and change seen in the Middle Irish period in particular will be an interesting challenge for either the "old" and "modern" FST/tagger. Dereza's (2018) Early Irish Lemmatiser will definitely have a role to play here, as it incorporates Middle Irish inflections given in eDIL; in

other words, we will (hopefully) arrive at the Early Irish eDIL headword. Adaptation of the Old Irish morphological FST to deal with "Middle Irishisms" is also a possibility. To properly deal with the verbal system of Middle Irish, a list of univerbated verb stems is necessary. Alternatively, or as a complementary approach, a list of prototonic compound stems from the Old Irish morphological FST can be extracted and combined with weak simple verb inflection. The latter will result in the generation of many non-existing or unattested "new" simple verb formations. Overgeneration, however, is not a problem from an analysis perspective (the grammatical analysis of an unattested form will never come up) and will enhance recognition scores.

Further possibilities include incorporating tags for language variety or linguistic features (such as for Middle Irish) on the lexical level of the (adapted version of the) Old Irish FST. Encoding this information will provide us with a way to augment morphological analysis with automatic textual dating. An application could be to establish the proportion of Middle Irish, as opposed to Old Irish, forms in an Early Irish text.

A more distant research prospect is the integration of POS taggers, databases, corpora and dictionaries into one lexical resource. Such a resource will hugely benefit scholars operating at the intersection of the Early and Modern Irish period, who now rely mainly on eDIL and Dinneen (1927), with no lexicographical facility that comprehensively spans the entire historical period. The author hopes to establish academic collaborations in the future to get a better grip on both the computational and linguistic challenges of his project.

Acknowledgement: This paper stems from research carried out during a Government of Ireland Postgraduate Scholarship (GOIPG/2017/1808) funded by the Irish Research Council. The author would also like to acknowledge the anonymous reviewer for helpful feedback and the editors for seeing this publication through.