Strategies in tracing linguistic variation in a corpus of Old Irish texts (CorPH)

David Stifter,[i] Fangzhe Qiu,[ii] Marco A. Aquino-López,[iii] Bernhard Bauer,[iv] Elliott Lash,[v] and Nora White[i]
[i] Maynooth University | [ii] University College Dublin | [iii] Centro de Investigación en Matemáticas | [iv] Karl-Franzens-Universität Graz | [v] Georg-August-Universität Göttingen

This article introduces Corpus PalaeoHibernicum (CorPH), a corpus currently consisting of 78 texts in Early Irish (c. 7th–10th cent.) created by the ERC-funded *Chronologicon Hibernicum* (*ChronHib*) project by bringing together pre-existing lexical and syntactic databases and adding further crucial texts from the period. In addition to being annotated for POS, morphological and syntactic information, another layer of annotation has been developed for CorPH – 'Variation Tagging', i.e. a tagset that numerically encodes synchronic language variation during the Early Irish period, thus allowing for much improved research on the chronological variation among the material. Another new pillar of studying linguistic variation is Bayesian Language Variation Analysis (BLaVA), in order to address the challenge that "not-so-big data" poses to statistical corpus methods. Instead of reflecting feature frequencies, BLaVA models language variation as probabilities of variation.

Keywords: Old Irish, diachronic variation, Chronologicon Hibernicum, Bayesian statistics, variation tagging

1. Introduction

Languages change constantly in all linguistic domains – phonology, morphology, syntax, and lexical use – and their graphic expressions are subject to fashions. Irish, a Celtic language spoken in Ireland, is in no way different. With a written history of more than 1,500 years, Irish is among the oldest attested languages in Europe. Because of its long textual tradition, its development through time is reflected in the huge amount of variation observable in the extant sources, i.e. texts in manuscripts from the 8th up to as late as the 17th and 18th century. The

https://doi.org/10.1075/ijcl.22018.sti | Published online: 20 September 2022
International Journal of Corpus Linguistics ISSN 1384-6655 | E-ISSN 1569-9811

European Research Council-funded project *Chronologicon Hibernicum* (hereafter *ChronHib*; 2015–2021) has studied the diachronic evolution of the early medieval Irish language, best known as Old Irish. This article presents the major challenges posed by extant Old Irish texts and introduces two methods developed in the *ChronHib* project to study synchronic and diachronic variation in the extant material, namely variation tagging and Bayesian language variation analysis.

2. Characteristics of Old Irish

Old Irish is commonly defined as the stage of the Irish language attested in the 8th–9th centuries CE, with the preceding 7th century constituting Archaic or Early Old Irish. Old Irish is followed by Middle Irish in the 10th–12th centuries, which in turn is followed by Modern Irish from c. 1200 onwards (see Stifter, 2009: 55). As a whole, the various pre-Modern stages of Irish are commonly referred to as Early Irish. Since the borders between these stages are fluid, to say the least, a pragmatically broad definition of Old Irish has been adopted in the *ChronHib* project, covering the period from c. 550–950 CE.

An example may show how drastic the changes even between Old and Middle Irish can be: in an early 9th-century manuscript now kept in Milan (Milan, Biblioteca Ambrosiana, MS C301 inf.), there are thousands of Old Irish glosses on a Latin commentary on the Psalms. One of these reads:

(1) *a* *n-as·mbeir-som* (Ml. 55b11 = S0006-3391)
when $^{\text{NAS}}$-PREVERB·$^{\text{NAS}}$say$_{\text{3sg.pres.ind.}}$-EMPHATIC$_{\text{3sg.}}$
("when he says")

This is the standard way of expressing this meaning in Old Irish, with two morphophonemic nasalisations (glossed as $^{\text{NAS}}$) prefixed to the preverb and to the stressed verbal root (following the raised dot), respectively. The same meaning is expressed in the *Passions and Homilies from the Leabhar Breac* from the 10th or 11th century as follows:

(2) *nuair adeir* *sé fēn* (Atkinson, 1887: l. 2688)
when say$_{\text{3sg.pres.ind.}}$ he CONTRASTIVE
("when he says")

The following changes have happened: the lexeme *nuair* "when" has replaced Old Irish *a* "when"; the compound verb *as·beir* has now amalgamated into the simple verb *adeir*, abandoning the distinction between the so-called deuterotonic and prototonic verb forms (McCone, 1997: 191–194); the nasalising mutation that signifies a temporal relative clause has also been lost in this construction; the independent pronoun *sé* "he" has emerged to mark the subject, which used to be subsumed in the zero-ending of the Old Irish verb form; finally, instead of the emphatic particle *-som* "he", contrastive *fēn* "himself, herself, etc." is used, which imbues the statement with a different pragmatic nuance and is not a linguistic change as such. The changes, which belong to different linguistic domains, happened at different times. They are in part the consequence of other changes and had themselves further knock-on effects on the grammatical system of medieval Irish. While it is easy to describe the difference by contrasting the surface realisation of the two phrases, a more precise chronological mapping of what is going on in between the two phases of Irish has proven much more elusive so far.

A large-scale chronological mapping of the development of a morphologically rich language such as Old Irish requires extensive linguistic annotations, in order to process the surface forms in a diachronic corpus as comparable data. Queries to retrieve information about linguistic variation and change rely on the accurate tagging of linguistic parameters. Once the data are annotated, it is possible to trace linguistic variation by defining the element that is liable to change ('variable'), retrieving all the different forms of this element ('variants') from the corpus, and comparing the variants.

In most cases, it is relatively straightforward to define a variable and to retrieve all the variants from the corpus with the help of annotations. For instance, one can define the semantic range of a given lexical item as the variable, and the actual meaning of this item in every instance in the corpus as the variant; one can then study the frequency and distribution patterns of the collocation of this lexical item in the corpus (Evert, 2008), measured against the timeline, which may reflect diachronic changes in its semantic range (e.g. Farr & O'Keeffe, 2002). In this case, the variants can only be categorised and hypothesised about *post hoc*, as one cannot possibly predict what meanings are assumed by relevant tokens before they are retrieved. Another approach is to first observe a finite set of variants, form a hypothesis about their relationships and then abstract from these variants the target variable, such as an inflectional category or a syntactic structure (e.g. Gries & Hilpert, 2010).

While some variants can be identified by querying a specific string and/or a tag, more complex selection of variants requires combining different annotation tags in a query. For example, in order to investigate the usage of English *help*, one can search for the combination of the lexical item HELP plus an infinitive verb with or without *to*, thus employing the lexical tags for retrieving HELP, the POS tags for allowing only verbs in the result and the morphological tags for identifying the infinitive. In this case, both the variable (the syntactic usage of *help*) and the variants (infinitive with vs. without *to*) are known before the search is conducted.

Not infrequently, retrieving all the relevant tokens requires a few rounds of trial and error to find the best balance between recall and precision. In their study on the dative shift alternation in English, Lehmann and Schneider (2012) test the automatic syntactic parser *Pro3Gres* (Schneider, 2008). They then decide to remove the lexical restriction and to add the preposition *for* to the restriction of verb-attached PP to achieve better inclusiveness of verbs (Lehmann & Schneider, 2012: 66–67). The final condition allows optimal recall of data, but also results in a lot of unwanted instances, or false positives, which are then manually excluded in the ensuing analysis on a case-to-case basis. Similarly, Hundt (2004) manually excludes constructions that are not progressive from her final sample pool of the instances of lemma BE + *-ing* gerunds in English. Schreier (2005), on the other hand, limits the lexical items that exhibit the phonological change #CCV- to #CV- in the history of English to a manually selected subset of data in order to ensure maximum accuracy of true positives in the returned result and the feasibility of analysis. Assessment of how over-inclusiveness or manual exclusion may affect the analysis remains a desideratum, but the point here is that there is no ready-made formula for how to retrieve all the tokens that represent variants of a certain variable from a corpus.

Where extant annotations in the corpora are inadequate for the research questions raised, individual studies often introduce additional tagging on corpus data. Even when the variants can be categorised by surface forms (e.g. *can't* and *cannot*), the categorisation actually constitutes an additional set of tags (i.e. type 1 = *can't*, type 2 = *cannot*), not to mention that more abstract categorisation is often needed, such as the types #CCV- against #CV- in Schreier (2005), or the 'object shift' word order in Old Norse which has been searched for by Rögnvaldsson and Helgadóttir (2011). In such cases, new tags have been added to the subsets of data that the scholars had retrieved for their individual research projects. However, such subsets of data with their additional annotations are usually published separately from the original corpora from which they were drawn. Some publications do not provide their full datasets at all, while other scholars share them through online repositories such as GitHub, but generally, such individually published datasets can be hard to locate and access. This may cause several problems:

i. Low replicability. It may be difficult for other researchers to independently replicate and verify the results, thus undermining the scientific reliability of the studies.

ii. Lack of transferability. The additional annotations that scholars created for their respective projects are seldom recycled for other projects and are thus underutilised.
iii. Repetitive work. Scholars may have to query, select and curate the same subset of data from large corpora every time the same variation is examined. Inconsistency may be introduced in the process as well.
iv. Limited research scale. The scattered nature of datasets and their inaccessibility hinder large scale quantitative research that looks at various types of variation at the same time. For instance, if one is to study the correlation between variation in the 3rd person and 2nd person singular verbal endings and the emergence of the *do*-auxiliary in Middle English, the amount of work would be significantly larger than when the datasets built for each variation type in previous research were readily available.

It is in the light of these problems that *ChronHib* has built its corpus of Old Irish texts and has designed an innovative strategy – tagging the variation within the corpus itself – to tackle them. This will be described in Section 5.

3. The corpus

A standard approach to tracking changes over time in historical linguistics is to operate with diachronic corpora, and then to compare the frequency of features at sequential periods through quantitative analysis (Hilpert & Gries, 2016: 36). Earlier large electronic corpora of medieval Irish texts are XML-marked-up corpora, such as the Corpus of Electronic Texts (CELT) (Färber, 2012–) and Thesaurus Linguae Hibernicae (TLH) (Kelly & Fogarty, 2006–2011), which are only tagged with textual and structural metadata, and are therefore inadequate for the purpose of detailed diachronic linguistic analysis. Fortunately, since the 2010s, a number of linguistically annotated corpora have emerged, drastically improving the possibilities for research on linguistic variation and changes in Old Irish. A list of these corpora and their contents and annotation schemes is provided in Lash et al. (2020: 2–3; see also Griffith et al., 2018: 11–18).

In terms of constructing a diachronic corpus, Old Irish texts pose a number of challenges. Firstly, the language has a highly complex morphosyntax, especially in the verbal system. A normal Old Irish verb encodes nine obligatory semantic, syntactic and morphological categories through formal means, to which two further facultative categories can be added (Stifter, 2009: 85–88): (i) person and (ii) number, (iii) tense and (iv) aspect, (v) mood, (vi) voice, (vii) relativity, (viii) deponentiality, (ix) dependency, (x) perspectivity, and (xi) object pronominality. Over 200 paradigmatic forms at least, with a large amount of allomorphic variation, can be formed for every verb. Each verb has to be annotated for each of these dimensions. For instance, the form *fiä* "you rested" is the independent 2sg. active (non-deponent) non-relative preterite (= punctual realis of the past) of *foäid* "to rest"; when the optional perspectival augment *ro-* is added, this regularly results in the perfect *·roä* "you have rested". On the other hand, *·fifam* "we will rest" is the regular dependent 1pl. active (non-deponent) future. Variable word boundaries and spelling variation aggravate the challenge posed by the rich morphology.

Secondly, despite being vast in extent and rich in genres, some of the most indispensable meta-information about early Irish texts is unknown or has not been established yet. For instance, the very number of extant medieval Irish texts is unknown, since there is no fully comprehensive catalogue that individually identifies every known Early Irish text. A provisional, project-internal list for *ChronHib*, based on earlier work, especially that of Ó Corráin (2017) and Hemprich (in preparation), contains 402 Old and Middle Irish prose texts, but the number of verse texts remains equally undetermined. More important than the absence of a reliable catalogue, however, is the lack of information for most texts about their author as well as time, place and intellectual environment of composition. This is compounded by the fact that the vast majority of early medieval texts are only found nowadays in late medieval or even early modern manuscripts, obviously having undergone centuries of transmission, during which younger linguistic forms may have crept in. Only for a comparatively small number of mostly short texts contained in manuscripts from the Old Irish period proper, catalogued in Bronner (2013), do their manuscripts provide at least a suitable *terminus ante* or *ad quem*.

In response to these challenges, *ChronHib* starts from the relatively few texts for which an approximate date is known, as determined predominantly by internal evidence and known manuscript dates, plus occasionally by known authors (see Toner & Han, 2019: 11–40, for methods of narrowing in on the unknown dates of texts). These are mostly texts preserved in contemporary manuscripts from 700–900 CE, i.e. mainly texts contained in *Thesaurus Palaeohibernicus* (Stokes & Strachan, 1901–1910). The majority of texts – 63 – are Irish glosses on Latin texts, but there are also four prose narratives, one gnomic text, one annalistic text, one long poem and six Latin texts that contain Irish personal and place names.

It is based on these texts that the Early Irish corpus of the *ChronHib* project has been built. Currently this corpus (Corpus Palaeohibernicum = CorPH; Stifter et al., 2021–) holds over 135,000 tokens from 78 Early Irish texts. Around 105,000 tokens are Early Irish, while 15,000 others belong to other languages (mostly Latin, but also some Old Norse, British or Anglo-Saxon proper names). Due to the nature of the preserved sources, the tokens are quite unevenly distributed across texts and across genres. Glosses from the Milan manuscript (Biblioteca Ambrosiana, MS C301 inf.) contribute almost 60,000 tokens, whereas some other texts have only a handful to a few dozen tokens.

The data entry has been informed by textual criticism. The researchers did not simply type in pre-existing editions of Early Irish text such as *Thesaurus Palaeohibernicus*, but – where available – the text was checked against high-resolution images of manuscripts, especially from the ISOS project (Dublin Institute for Advanced Studies, 1999). In this way it has been possible to emend manifest errors in the received editions. This procedure has led to revised texts in CorPH that effectively supersede pre-existing editions and it transcends traditional editorial paradigms in the discipline. CorPH fulfils two separate functions at the same time, namely creating a searchable corpus *and* providing reference texts that conform to up-to-date philological standards.

A detailed description of the structure of CorPH and the collection and processing of data can be found in Qiu et al. (2018) and Qiu and Stifter (2020). Briefly speaking, CorPH is a lexicographically oriented corpus. It has inherited a large amount of data from previous projects that were concerned with creating specialised lexica of individual Old and Middle Irish texts (see Griffith et al., 2018: 11–18), namely Griffith and Stifter (2013), Bauer (2015), Barrett (2017), and Bauer et al. (2017). For some of the data, it has utilised the Parsed Old and Middle Irish Corpus (Lash 2014), a syntactically parsed treebank of early Irish texts. The rest of the data was tagged as part of the project work. The starting point was Griffith and Stifter's database (2013), the basic structure of which was kept for later lexical databases. These databases were mostly built using the *FileMaker Pro* software versions 8–14 (Claris International Inc., 2006–15), which has become obsolete for this type of application in the meantime. Much effort had to be put into harmonising the different annotation schemes used in these databases and into transferring the data into an updated, more universal format. The process has been highly time-consuming but at the end we have managed to harmonise all the data in a new database using JavaScript, SQL, ExpressJS, Angular 9+ and NodeJS stack hosted on a Microsoft Azure Cloud server managed by Maynooth University. The corpus is accessible to the public via https://chronhib.maynoothuniversity.ie/.

As a consequence of its focus on individual tokens, the architecture of CorPH is constructed as follows. CorPH is built as a relational database that consists of four major tables: 'Morphology', 'Lemmata', 'Text', and 'Sentence'. The 'Morphology' table constitutes the core annotated corpus and includes lexical, morphological and syntactic information, including a full grammatical analysis for each token of early Irish. Information pertaining to lexemes rather than concrete tokens, such as POS, meaning, and etymology, is stored in the 'Lemmata' table. Metadata about texts, such as their date, scribe, and bibliographical information, are found in the separate 'Text' table. The metadata about individual textual units, whether they be sentences, lines of poem, or glosses, are in yet another table called 'Sentence', which also includes Latin texts associated with individual glosses, translations of the Irish textual unit, location in the manuscripts, editorial comments, etc.

CorPH tokenises Old Irish texts into the smallest *lexically* analysable units, called 'morphs'. This idiosyncratic term (not to be confused with Haspelmath's, 2020, very different use of 'morph') is driven by practical considerations. *A Dictionary of Linguistics and Phonetics* (Crystal, 2008: 313) defines a morpheme as "the minimal distinctive unit of grammar, and the central concern of morphology". Therefore a morpheme, such as an inflectional ending, is not necessarily a lexically meaningful unit, though it may of course have a *grammatical* function. However, many elements that correspond semantically and etymologically to lexical items, i.e. 'words', in Standard Average European languages, e.g. conjunctions or object pronouns, can only occur as bound morphemes in Old Irish. Even though etymologically these elements all go back to independent words in the prehistory of Irish, synchronically they have to be incorporated into an accentual domain such as the verbal complex or are cliticised to nouns. Occasionally, they even turn into independent words again over the course of time. For practical reasons and to adequately represent the morpho-syntactic structure of Old Irish, especially in the verbal system (see above), it is therefore desirable to have a way of speaking about such items, which occupy a grey area between morphemes and 'words'. On the other hand, inflectional morphology, such as case, tense, or subject-agremeent marking, has not been treated in a similar way because its diachronic behaviour is very different from those lexical items mentioned above. Additionally, it would not be possible to capture Old Irish inflectional morphology with a manageable system of annotation due to its enormous complexity.

In CorPH, facultative elements such as conjunctions or object pronouns, which have been attracted into the orbit of the verbal system, are not understood as part of verbal morphology, but are treated rather as separate tokens. These functional elements usually take the form of overt morphemes, but through post-morphological processes some of them can be deleted again on the surface. For instance, the addition of the perspectival augment *ad-* to *con·toil* "you slept" results in *con·atoil* "you have slept" < **kom-ad-tol-es-*. Adding the main clause negative particle *ní·* turns this into *ní·comtoil* "you have not slept", which still contains *ad-*, although it has been deleted from the surface through synchronic morphophonetic processes. CorPH includes such reduced or deleted elements as tokens, i.e. as 'morphs', in the corpus.

The bundle of elements that is drawn into the accentual domain of the verb is traditionally called 'verbal complex'. Consider the example *amail dund·rigensat* “as they had done it” from Textual Unit S0006-4331. This has several bound morphemes as infixes: -*Ø*- “nasalising relative marker”, -*nd* “infixed pronoun 3sg. neuter”, *ri*- “perspectival augment”. They each receive morphs of their own, as well as the lexical preverb *du*, so that this verbal complex is tokenised as shown in Table 1.

Table 1. Tokenisation in CorPH: *amail dund·rigensat* “as they had done it”

Morph	Analysis	Lemma	POS	Classification	Meaning
amail		*amail 2*	conjunction		as, like; translating Latin *ut*
du		*de·, dí·*	particle_preverb		aspectual and lexicalized meanings
Ø		*nas.rel.particle*	particle_relative		nasalising relative particle
nd·	C.	*3sg.neut.inf.pron.*	pronoun_infixed		it
·ri		*ro· 2*	particle_augment		with pres. ind. and subj. with perfective sense; to express potentiality or possibility, with all forms of the verb; perfective augment
dund·rigensat	augm.3pl.pret.	*do·gní*	verb	H2	to do, to make

As can be seen, CorPH not only contains lemmata that correspond to traditional dictionary headwords, but also abstract entities such as *3sg.neut.inf.pron.*, whose surface representation can be -*a*, -*t*, -*did*, *Ø* or, as in this example, -*nd*. There are practical limits to what can be annotated: while the rather broad present-stem formation of verbs is encoded in the ‘Classification’, the intricately complicated formation of other temporal and modal stems is not captured in this system.

4. *Corphusator*

As a technical consequence of the annotation procedure in *ChronHib*, the source texts are not entered in CorPH in simple linear fashion with annotation of one item after the other, but the texts are rather broken down into isolated lexical or morphemic chunks. The chosen type of annotation leads to a relatively large amount of item redundancy which precludes a bidirectional one-to-one mapping between original text and its representation within CorPH. For instance, a simple concatenation of the single morphs of S0006-10 *co du·fobither* "that it be cut down" would be: *co*, *du·*, *·fo*, *du·fobither*. In this example the redundancy is caused by the two preverbs *do·* and *fo·*, which occur both on their own but also within the verbal form in CorPH. To give an example of a compound noun, S0006-6169 *óinmenmnaige* "being of one mind, concord" is included in the data-base with entries for *óin*, *menmnaige* and *óinmenmnaige*. Therefore, it is not possible to straightforwardly export textual corpus files from the lexical database. For a small field like Early Irish Studies, where textual philology and linguistics are intimately intertwined, the availability of a textual corpus in the traditional sense is, however, desirable; creating a database for lexical investigation that does not at the same time serve as a textual corpus would be a waste of effort. The limited funding that is typically awarded to the field in its entirety needs to be used as efficiently as possible.

However, an extra step is necessary to convert CorPH into a textual corpus of Early Irish in the traditional sense. For this purpose, Bernhard Bauer created a special software, *Corphusator* (https://github.com/BernBa/Corphusator), that automatically removes the project-specific redundancies and converts the lexical entries into tagged .txt files (see Bauer, in preparation). These files can be freely downloaded from the project's Github account (https://github.com/chronhib-MU/Chronhib-Website/tree/master/client/src/assets/docs) in three different formats: (i) POS-tagged, (ii) morphologically tagged, or (iii) POS- and morphological tags combined.

5. Variation tagging

CorPH forms the basis on which we can examine these texts for a large set of linguistic features that are known to have undergone change in the course of Early Irish. The entirety of the specific values that those features take for a single text create what can be called 'synchronic linguistic profiles' within the continuum of multivariate change. By comparing these synchronic profiles, we are able to delineate the diachronic trends of development of Old Irish. The selection of the linguistic features to be examined is of paramount importance, but equally important is how to keep track of a large number of features in multivariate comparison, and how to avoid the several problems that we identified at the end of Section 2. A solution is proposed in the following.

By trawling through scholarly literature, especially treatises on the diachrony of the Irish language (e.g. Thurneysen, 1946; McCone, 1996; Schumacher, 2004) and critical editions of texts (e.g. textual editions in the major series such as those published by the Dublin Institute for Advanced Studies or the Irish Texts Society, and editions in journals such as *Ériu*, etc.), we have compiled a list of more than 300 changes in the areas of phonology, morphology, orthography, syntax and lexicon that are known to have happened within the investigated period. These constitute the range of variables on the basis of which we can profile the externally dated texts and depict linguistic change throughout the period. Moreover, we are in an advantageous position that for the majority of these variables, the variants are finite, and their relative chronology is clear. Below in Table 2 are a few examples of historical changes in Old Irish. Each change is assigned a unique ID and is categorised according to the grammatical area it belongs to (phonology, morphology, syntax, as well as orthography), and it receives a brief verbal description ('X > Y' means 'X becomes Y').

Table 2. Samples of linguistic variation in Old Irish

Description of change	Category	ID assigned to change
Stressed /au/ from u-infection becomes /u/ or /o/, e.g. *maug* > *mug*; *aub* > *ob*	1. Phonological	PH019
Use of double vowels to represent long vowels. e.g. *maar* for *már*	2. Orthographical	OR015
In strong verb inflection, present indicative/present subjunctive/ imperative singular passive ending *-a(i)r* > *-tha(i)r*	3. Morphological	MO011
Reduplicated suffixless preterite inflection was replaced by s-preterite inflection, e.g. *·lil* > *·len* "followed", *·sefainn* > *·seinn* "pursued"	3. Morphological	MO014
Accusative replaces dative as complement to comparative adjective	4. Syntactical	SY007

A dated text belongs to a fixed period of time, so if we can count the tokens for each variant of the same variable in the text, the percentage of variants can be regarded as a proxy for the profile of this variable at this particular time. Measurements of different variables constitute the linguistic profile of the text, and profiles of multiple contemporary texts provide a profile of the language at the time. The data of the linguistic profiles at different times will allow the calculation of the rate of change and the correlation between changes. With sufficient data, it would ideally even be possible to predict the profile of the language at a given time, but the practical limits of the attested materials only allow approximations.

The problem is how to retrieve all tokens that contain the desired variants of a specific variable from the corpus. Variants of some variables can be retrieved by means of the existing grammatical tags. For example, in order to find tokens that display the two variants in SY007, as described in Table 2 above, one needs to search for all NP complements to tokens that are tagged morphologically as *comp.* (standing for the comparative grade of the adjective). The heads of these NPs have morphological tags either as *acc.* (accusative) or *dat.* (dative). By observing the distribution of these two tags, one would be able to describe over which period this change occurred and in what way the innovative construction became dominant. To find tokens relevant to change MO011, a query with the following conditions is created. The 'Analysis' field of the 'Morphology' table should have one of the following three tags, *3sg.pres.ind.pass.*, *3sg.pres.subj.pass.* and *3sg.impv.pass.* At the same time, relevant tokens should have *S%* (standing for strong verb) as the tag in the 'Classification' field in the 'Lemmata' table. This query shows all the strong verbs in the present indicative, present subjunctive and imperative singular passive, which are susceptible to the change MO011.

However, finding the variants for some other variables is not so straightforward. At first glance, OR015 can be captured quite easily by searching for all strings that consist of two identical vowel letters in a token. However, this will return many false positives that contain two identical vowels separated by a hiatus, such as *biid* "of food", and one has to manually exclude these tokens. Similarly, for PH019 it is impossible to simply search for tokens that contain the strings *au*, *u* and *o*, as the result will be inundated by irrelevant tokens. There is simply no tag by which we can tell the computer to identify a token that would have contained an *u*-infected vowel /au/ in early Old Irish. That judgment relies heavily on expert knowledge of etymology and historical phonology. Likewise, if we wish to query for the change MO014 by which the reduplicated suffixless preterite inflection was replaced by the s-preterite inflection in the verbal paradigm, there is no tag specifically dedicated to the preterite stem formation, nor is there a unified criterion by which the machine could tell the inflectional class from a token's surface form.

In order to study changes like PH019 and MO014, manual selection of relevant tokens from the corpus seems inevitable. One can, of course, reduce the workload by pre-selecting a small subset of better attested lexical items that are known to contain the variants, observe the distributions of variants in this subset, and assume that sample is representative of all tokens affected by that change. For example, to investigate the change MO014, one may extract all preterite forms of the verbs *gonaid* "wounds", *ar·cain* "recites" and *maidid* "breaks", which originally have reduplicated suffixless preterite in Old Irish, and treat their developments as representative of the reduplicated suffixless preterite in general. However, because the size of CorPH is relatively small, and since other factors such as the phonological shape of the verbal stem may have affected the change, it is preferrable not to limit the study to selected lexical items but to manually pick out all tokens that contain the variants in question from the entire corpus. In the case of MO014, this means that we would have to go through all preterite verbal forms in the corpus and manually retrieve those that originally had a reduplicated suffixless preterite, based on our specialised knowledge. Within this subset of MO014-relevant tokens, we would then categorise the tokens into (i) those showing reduplicated suffixless inflection, (ii) those showing s-preterite inflection, and (iii) other forms, including unclear cases, therefore creating new tags on the data.

This means that every time a change is investigated, a subset of the data is selected and annotated for that specific purpose, often involving heavy manual curation. However, since the goal of the *ChronHib* project is to combine and compare different variables to create textual or temporal profiles, the retrieved subsets have to be reused and compared over and over again. As mentioned above, we profile individual texts T_1, T_2,…, T_n repeatedly on the same variables V_1, V_2,…, V_n, before we combine the synchronic profiles of texts into a diachronic profile of Old Irish. Accordingly, it is more practical to annotate the tokens directly in the corpus with the relevant variant and variable, so that every time a variant or a variable needs to be examined, there is no need to run that query again on the whole corpus. In essence, we are proposing a standardised variation tagset for annotating the tokens.

The tagset contains a list of variables. For each variable, the possible variants are given the number codes 1, 2, …, n. The sample list in Table 3 is thus updated as follows:

Table 3. Updated list of variation types

ID	Category	Description
PH019	1. Phonological	stressed /au/ from u-infection appears as 1. *au* 2. *u* 3. *o*
OR015	2. Orthographical	long vowel is represented by 1. double vowel letters 2. single vowel letter
MO011	3. Morphological	in strong verb inflection, present indicative/present subjunctive/ imperative singular passive ending is 1. *-a(i)r* 2. *-tha(i)r*
MO014	3. Morphological	Originally reduplicated suffixless preterite form shows 1. reduplicated suffixless inflection 2. s-preterite inflection
SY007	4. Syntactical	complement to comparative adjective is in 1. accusative 2. dative

The tokens in CorPH are tagged with the ID of the variable and the number code of the variant, whenever the variable is applicable to that token. The number 0 is reserved for indeterminable cases, where it is uncertain which variant the token shows. There is no limit on how many variation tags a token can have. Examples are given in Table 4.

Table 4. Examples of variation tagging

Morph	Variation tag	Description
Culand	PH019.2	stressed /au/ from u-infection is 1. *au* 2. *u* 3. *o*
	PH057.2	original /nd/ in places other than proclitics is 1.*nd* 2.*nn*.
	PH029.2	posttonic, non-final short vowel is 1. unchanged 2. schwa
Feradach	PH029.0	posttonic, non-final short vowel is 1. unchanged 2. schwa
das	MO072.2	infixed pronoun shows 1. conservative forms 2. innovative forms

In the token *Culand* "Culann (a personal name)", the annotator notices that the etymology is **kaluno-*, and deduces that the first syllable should have been *au* by *u*-infection in early Old Irish. Therefore PH019 applies in this case and the variant is *u*, as it appears in the surface form. The original *n* in **kaluno-* became *nn* by the so-called 'MacNeill's law' in early Old Irish, and later, when the change PH057 occurred, original *nd* also became *nn*. Since the token *Culand* does not have an original *nd* in its etymon, the *nd* in the token can only be hypercorrect, which means that by the time of this token, PH057 had already occurred and original *nd* and *nn* (<**n*) were now confused, so that, in line with a much wider orthographic tendency, the scribe tried to "restore" the *nn* in the contemporary language to what he erroneously believed to be the "older" *nd*. The second syllable in the token should have been *o* in early Old Irish (i.e. **Caulonn*), but in this token, the posttonic, non-final short *o* is spelled *a*, suggesting that PH029 has occurred and the vowel is now a schwa. On the other hand, the *a* in the second syllable of *Feradach* "Feradach (a personal name)" can represent the vowel /a/ or schwa, which are undistinguishable in this position by Irish orthography. If we can prove that this vowel should have originally been /e/, then it is clear that the letter *a* must represent a schwa, as there is no phonological process by which /e/ could become /a/ in this position. But without the help of etymology for this word, we are not able to judge what phonetic value this vowel letter represents, and consequently we tag it with the value 0. For the token *das* "them", which is an innovative form arising as a hybrid of two different types of infixed pronoun (that of -*da* and -*s*, respectively), there is no doubt that it should be assigned the tag of MO072.2.

Because of its benefits, variation tagging is thus an integral part of the CorPH annotation scheme alongside the more conventional linguistic tags mentioned in Section 2 above. It feeds into and hugely facilitates the statistical analysis discussed in Section 6.

6. Bayesian language variation analysis

The way the mark-up for POS, morphological and variational information has been done allows for very detailed quantitative analysis of the data, not only for frequencies of lexical usage or the occurrence of surface morphemes and constructions, but also for the quantification of more complex linguistic variation and change over time. Here, finally, the dearth of secure and precise dates for the texts, mentioned earlier, becomes a major factor. Starting from the relatively few texts of which a date is known or for which an approximate date can be estimated from internal and external evidence, the distribution of variants of particular variables can be arranged in a chronological sequence. This can be done for single or for multiple variables, up to the more than 300 types of synchronic variation that have been identified for Early Irish so far. As mentioned above, the individual values for all the synchronic variation as manifested in a specific text "package" constitute its variational-linguistic profile.

Ideally, tracking the differences in values for specific variables would result in neat graphs that show the gradual change of individual features over time. In actual practice, the development may present itself rather "dirty": depending on the feature under scrutiny, there may be more or less data available, and often it falls below the level of statistical significance. Depending on the very unevenly balanced size of the sources used, the evidence is much more significant for some periods than for others. And finally, frequencies of particular features may not only vary according to time, but also according to fuzzy and non-controllable factors such as genre, stylistic preferences, and place. If one were to track the mere frequencies of certain features, the emerging picture might not allow for discovering a trend at all. Finally, even the assumed dates are fuzzy in that exact calendar years are known only for an exceedingly rare number of texts. In the majority of cases, the dated texts have rather been assigned to a probable range of years or decades, which can be called the 'dating interval'. For instance, the Milan Glosses are found in a manuscript that can be dated palaeographically and codicologically to the first half of the 9th century (Bronner, 2013: 27). Since the Old Irish glosses in the manuscript are manifestly copied (the numerous, typical copying errors are a tell-tale signal of this), their original composition must be older than the extant manuscript. In CorPH, a dating interval of 780×810 has been judged to reflect the most probable date.

This means that in practical terms the data is characterised by a great amount of uncertainty. It is here that another powerful method comes in, namely Bayesian statistics. The use of Bayesian statistics enables researchers to make probabilistic statements about research questions, which is a more realistic reflection of the uncertainties that are inherent in the data. The use of Bayesian methods has seen an upsurge in historical linguistics in recent years. The most common application is for establishing phylo-genetic relationships within language families and for dating the bifurcations in language trees (see the recent examples of Rama & Wichmann, 2020, or Sagart et al., 2019). The use of Bayesian methods to date linguistic developments within a historically attested language has so far only been attempted by Hellwig (2019) and (2020) for Vedic Sanskrit, independently from and parallel to the work in *ChronHib*. Hellwig's main model is a Bayesian Mixture model which infers ages from linguistic characteristics, and assumes that every word comes from the same age (the age of one of the books of the Ṛgveda).

Bayesian Language Variation Analysis (hereafter BLaVA), the statistical method developed in the *ChronHib* project by Marco Aquino-López, differs from Hellwig's approach in that it can be described as a Bayesian logistic regression over the variations of a single word, form or feature. Instead of tracking the absolute frequencies across the texts in the corpus, BLaVA allows the user to model the probability of variation changes over the period under observation; this is especially useful when we want to date a shift in the use of a particular variation. This is done by inferring probabilities for the usage of a variant over another for a period of time. Because each occurrence of a form can come from different periods of time (even if found in the same document), it was necessary to allow the model the freedom to vary each occurrence within the period of the suspected date of the text. This achieves two main goals: it takes into consideration the uncertainty of the dating of each individual occurrence of a form, and it allows us to observe the posterior age of each occurrence.

The logistic regression part of the model allows us to infer the probability of the occurrence of a variant at any particular time within the studied period, i.e. how likely it is to observe a particular variation of a form at any given time. Each parameter of the logistic regression is assigned a prior distribution. Given that we do not have initial suspicions as to which variant is predominant, we provide the model with prior distributions which do not favour one variant over another (i.e. normal distribution centre at 0). In order to perform the logistic regression, one variant is chosen as the base variant and is assigned the value 1, and the resulting model will represent the probability of observing this variant. When the other variant appears, 0 is assigned. In order to obtain the probability of this variant we can calculate it as $p_0=1-p_1(t)$. The probability p_0 is calculated as,

$$p(t) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 t)}}$$

where *p(t)* represents the probability of using the base variant at time t and β_o and $\beta_1 t$ are the parameters of the model. Because each variant of a form can occur within the chosen period, each occurrence was assigned a normal distribution as the prior distribution; the mean is set at the mid-point of the suspected dating interval and the standard deviation defined such that 95% of the probability of the distribution lies within the above-mentioned interval. By defining the model in *R* ver. 4.0 (R Core Team, 2020) and using R2JAGS ver. 0.6–1 (Su & Yajima, 2020), a sample of the posterior distribution of each parameter can be obtained.

Once a sample of the posterior distribution of each parameter is obtained, we can use the logistic link function to calculate the posterior probability of the occurrence of one of the variations at any time within the period under study. It is important to keep in mind that in this way we are able to track the probability, not the frequency, of variation over time and when the probability of one variant overtakes another.

An example from CorPH will illustrate the method, namely the variation between the two allomorphs of the proximal demonstrative particle "this": *-so* and *-se*. This particle is enclitic to nominals. There are two possible relevant phonological contexts: either the noun preceding the particle ends in a non-palatalised 'neutral' sound, i.e. a non-palatalised consonant or the back vowels *a*, *o*, *u*, or in a palatalised consonant or the front vowels *e* and *i*. The phonological opposition between neutral and palatalised contexts is crucial to Old Irish grammar and extends to all consonants (Stifter, 2009: 62). Examples for all four possible commutations are neutral/back context + *-so*: *andubso* "this ink" (from the Textual Unit S0006-66), palatal/front context + *-so*: *indfirso* "of this man" (S0006-710), palatal/front context + *-se*: *innafaithsinese* "of this prophecy" (S0006-845), neutral/back context + *-se*: *infectse* "this time" (Lash, 2014: lc.208).

Figure 1 shows the distribution of the variant (i.e. allomorph) *-se* across time and in the two phonological contexts (sample size: 267; 198 examples from CorPH, plus 69 examples from additional sources, especially Lash, 2014 and Kavanagh, 2001). The x-axis shows the year, the y-axis shows the probability of *-se* to occur.

Figure 1 cannot be interpreted to reflect directly the relative frequency of the variants. Instead, the area delimited by the red graphs indicates the probability of finding *-se* after neutral sounds, the blue ones that after palatalised sounds. The outer dashed lines show the 95% confidence intervals; the middle lines represent

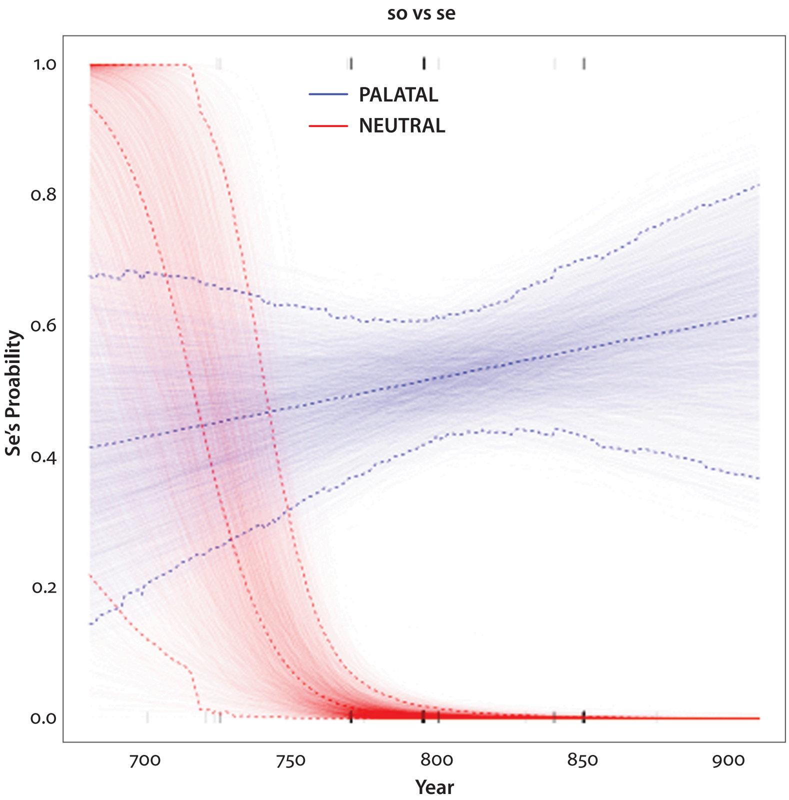

Figure 1. BLaVA of the *-so*/*-se* variation

the mean of the probability of *-se* to occur. Black lines at 0 and 1 are the mean of the posterior distributions of each of the occurrences of *-se* (at 1) and *-so* (at 0). The probability of *-so* is the inverse of that of *-se*. From this figure it can be inferred that *-se* was most likely never predominant in neutral contexts, but that it completely died out there by the 800s. The enormous spread around 700 only means that for this period the probability for the occurrence of the variant could be anywhere between 25%–100%. The paucity of available data does not allow us to be more precise than that. On the other hand, in the palatalised context, the figure reveals a slow but clear increase of the probability of use of *-se* compared with *-so*, from c. 40% to 60% over the three centuries.

If we are interested in a particular year, we can look at the posterior distribution of the probability of the variant at that time. To illustrate this, Figure 2 shows the posterior probability of the occurrence of *-se* taken at two cross-sections in Figure 1, the years 750 and 760. The graphs on top correspond to the probability demarcated by red lines in Figure 1, the ones on the bottom to the blue lines. In

750 the probability intervals still overlap largely which means we cannot guarantee that the behaviour of *-se* differed between palatalised and neutral contexts. By 760, however, the intervals are completely separated, which means we can be confident that the behaviour in both categories was different. The cross-sections reveal more about the internal probability distribution for each year, which cannot be illustrated in the long-term graph, but they lack the dynamic information contained in the latter. The results of the Bayesian language variation analysis of *-se* and *-so* are wholly new and contradict what has been said about their distribution in previous literature (e.g. Uhlich, 2018).

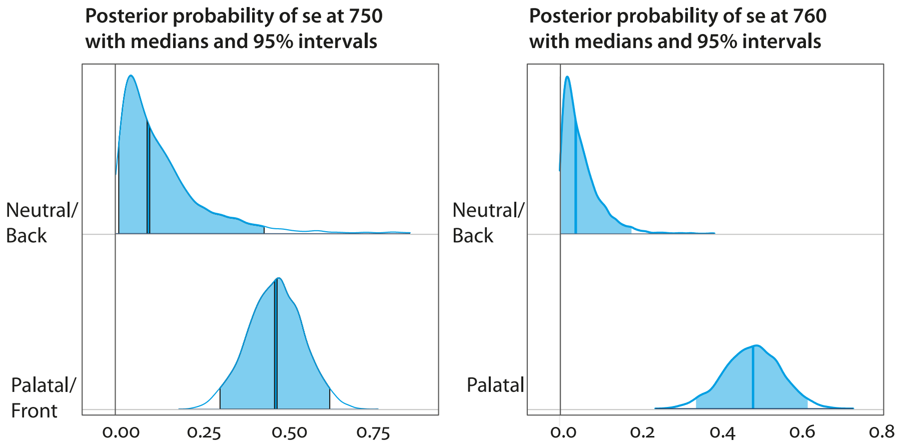

Figure 2. Posterior probability of the occurrence of *-se* at 750 and 760, indicated by the x-axis

7. Advantages and benefits of the methods

Two new methods of studying historical language data have been developed and tested in the *Chronologicon Hibernicum* project, namely variation tagging and Bayesian language variation analysis.

Variation tagging adds a whole new type of annotation to tokens in a historical corpus. It serves to encode synchronic variation and diachronic change directly on a token and it thereby creates a shortcut that avoids having to conduct complicated searches on the corpus. It also makes such data transferable and reusable. It allows for the inclusion of annotation for features that could not or not easily be captured by searches for surface forms in corpora with conventional annotation schemes, therefore it has a great potential in documenting more complex types of linguistic variation. Different researchers can study the same variation by a simple query for the relevant tag without having to duplicate the work. Any token that is relevant to a specific variation can be easily retrieved and reused for different purposes. In addition, variation tagging facilitates comparison between different linguistic variations, say if one wishes to find out whether there is any statistical correlation between the phonological change from *nd* to *nn* and the similar change from *ld* to *ll* in Old Irish. Furthermore, this tagset standardises the description of linguistic variation, and is open to newly discovered variants, which can simply be assigned a consecutive number in the relevant variable. It is likely that the number of tags will increase in future work. The tagset used in CorPH is only specific to Old Irish, but it is easy to define a similar variational tagset for any other language.

Bayesian Language Variation Analysis (BLaVA) allows us to model language variation as probabilities of the occurrence of variants of a variable. This can be useful where data is too sparse, fuzzy or messy to allow straightforward frequentative analyses, as is commonly the case in corpora of underrepresented, historical languages. BLaVA works with any kind of annotation that allows us to compare different distributions of values, but employed in conjunction with variation tagging it is a powerful tool to model the historical development of languages such as Old Irish and to make progress in a more precise understanding of their chronology.

Both methods are still only in the initial phases of being applied and tested. Old Irish, with its particularly difficult historical phonology and its complex morphology and morphophonemics, is a good testing ground to sound out the potential of the two methods.

8. Challenges and desiderata

There have been numerous challenges and practical difficulties encountered in developing and applying the two methods of variation tagging and Bayesian language variation analysis introduced in the foregoing sections. As for the corpus CorPH itself, future work will have to involve greater diversification in several respects. Where places of origin are known or can be inferred from extra-linguistic criteria, greater regional diversity will have to be aimed at in order to potentially uncover diatopically distinctive features. At the moment, CorPH has a strong emphasis on the north-east of Ireland. It will also be necessary to extend CorPH to a wider range of genres, and to boldly go beyond Old Irish, namely into Middle Irish, but also into the fragmentary remains of the earlier stages of the language called Primitive Irish (4th–6th centuries CE). As for core Old Irish, some important early texts are missing: the Würzburg Glosses, the Cambrai Homily, and the Stowe Treatise on the Mass, as well as early examples of poetry.

The main practical challenge in building the corpus is the large amount of manual labour required for the complex tagging with respect to POS, morphology, syntax and variation. Since a large number of subtle rules have to be observed in the annotation process, maintaining perfect uniformity across the presentation of the material is a challenge, not only across several contributors, but even for a single contributor. In the future, software support needs to be developed in order to speed up the process. It is essential that the tags in CorPH are as consistent and as accurate as possible within the limits of our current understanding of the diachronic and synchronic grammar of Old Irish. Given the complexities involved (a morphologically complex language, occasionally corrupt manuscript transmission, etc.), full automatisation of the tagging process will not be possible. Because of the ambiguities and many homonymies in medieval Irish, the human interface and human understanding of texts will remain essential as a corrective force for the foreseeable future.

Neither Bayesian language variation analysis nor variation tagging as a method as such are specifically tailored towards Old Irish. They can be applied to any historical language of restricted documentation that shows sufficient phonological and morphological variation within its corpus. However, in the case of variation tagging, the concrete tags have to be determined specifically for each target language. The currently used tags are only applicable to medieval Irish in its various manifestations.

Funding

This project has received funding from the *European Research Council* (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 647351).

Acknowledgements

This article was written as part of the research project *Chronologicon Hibernicum* (https://www.maynoothuniversity.ie/chronologiconhibernicum). Our thanks go to all our colleagues in the *ChronHib* project, Francesco Felici, Ellen Ganly, Tianbo Ji, Truc Ha Nguyen, Godstime Osarobo, for their work and support, and to two anonymous reviewers.

References

Atkinson, R. (1887). *The Passions and the Homilies from Leabhar Breac.* Royal Irish Academy.

Barrett, S. (2017). *A Study of the Lexicon of the Poems of Blathmac Son of Cú Brettan*. [Doctoral dissertation, Maynooth University]. MURAL – Maynooth University Research Archive Library. https://mural.maynoothuniversity.ie/10042/

Bauer, B. (2015). *The online database of the Old Irish Priscian Glosses*. http://www.univie.ac.at/indogermanistik/priscian/

Bauer, B. (in preparation). *Corpus Palaeohibernicum (CorPH): From an Early Irish lexical database to a text-based corpus using Python*.

Bauer, B., Hofman, R., & Moran, P. (2017). *St Gall Priscian Glosses* (Version 2.0). http://www.stgallpriscian.ie

Bronner, D. (2013). *Verzeichnis altirischer Quellen* [Directory of Old Irish Sources]. Philipps Universität Marburg.

Claris International Inc. (2006–15). *FileMaker Pro 8–14*. [Computer Software]. https://www.claris.com/filemaker/

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics*. (6th ed.). Blackwell. https://doi.org/10.1002/9781444302776

Dublin Institute for Advanced Studies. (2004–). *Irish Script on Screen*. https://www.isos.dias.ie/

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (pp. 1212–1248). Mouton de Gruyter.

Färber, B. (2012–). *CELT: Corpus of Electronic Texts*. http://celt.ucc.ie/

Farr, F., & O'Keeffe, A. (2002). *Would* as a hedging device in an Irish context: An intra-varietal comparison of institutionalised spoken interaction. In S. M. Fitzmaurice, D. Biber, & R. Reppen (Eds.), *Using Corpora to Explore Linguistic Variation* (pp. 25–48). John Benjamins. https://doi.org/10.1075/scl.9.04far

Gries, S. Th., & Hilpert, M. (2010). Modeling diachronic change in the third person singular: A multifactorial, verb- and author-specific exploratory approach. *English Language and Linguistics*, *14*(3), 293–320. https://doi.org/10.1017/S1360674310000092

Griffith, A., & Stifter, D. (2013). *Dictionary and Database of the Old Irish Glosses in the Milan MS Ambr. C301 inf.* https://indogermanistik.univie.ac.at/milan-glosses/

Griffith, A., Stifter, D., & Toner, G. (2018). Early Irish lexicography – A research survey. *Kratylos*, *63*, 1–28. https://doi.org/10.29091/KRATYLOS/2018/1/1

Haspelmath, A. (2020). The morph as a minimal linguistic form. *Morphology*, *30*, 117–134. https://doi.org/10.1007/s11525-020-09355-5

Hellwig, O. (2019). Dating Sanskrit texts using linguistic features and neural networks. *Indogermanische Forschungen*, *124*, 1–47. https://doi.org/10.1515/if-2019-0001

Hellwig, O. (2020). Dating and stratifying a historical corpus with a Bayesian mixture model. In R. Sprugnoli & M. Passarotti (Eds.), *Proceedings of the LREC 2020 1st Workshop on Language Technologies for Historical and Ancient Languages* (LT4HALA 2020) (pp. 1–10). European Language Resources Association. https://aclanthology.org/2020.lt4hala-1.1.pdf

Hemprich, G. (in preparation). *Catalogue of Medieval Irish Literature*.

Hilpert, M., & Gries, S. Th. (2016). Quantitative approaches to diachronic corpus linguistics. In M. Kytö & P. Pahta (Eds.), *The Cambridge Handbook of English Historical Linguistics* (pp. 36–53). Cambridge University Press. https://doi.org/10.1017/CBO9781139600231.003

Hundt, M. (2004). Animacy, agentivity, and the spread of the progressive in Modern English. *English Language & Linguistics*, *8*(1), 47–69. https://doi.org/10.1017/S1360674304001248

Kavanagh, S. (2001). *A Lexicon of the Old Irish Glosses in the Würzburg Manuscript of the Epistles of St. Paul* (D. S. Wodtko, Ed.). Österreichische Akademie der Wissenschaften.

Kelly, P., & Fogarty, H. (2006–2011). *Thesaurus Linguae Hibernicae.* https://www.ucd.ie/tlh/index.html

Lash, E. (2014). *The Parsed Old and Middle Irish Corpus (POMIC)* (version 0.1). https://www.dias.ie/celt/celtpublications-2/celt-the-parsed-old-and-middle-irish-corpus-pomic/

Lash, E., Qiu, F., & Stifter, D. (2020). Introduction: Celtic studies and corpus linguistics. In E. Lash, F. Qiu, & D. Stifter (Eds.), *Morphosyntactic Variation in Medieval Celtic Languages: Corpus-based Approaches* (pp. 1–12). De Gruyter Mouton. https://doi.org/10.1515/9783110680744-001

Lehmann, H. M., & Schneider, G. (2012). Syntactic variation and lexical preference in the dative-shift alternation. In J. Mukherjee & M. Huber (Eds.), *Corpus Linguistics and Variation in English: Theory and Description* (pp. 65–75). Rodopi.

McCone, K. (1996). *Towards a Relative Chronology of Ancient and Medieval Celtic Sound Change.* Maynooth.

McCone, K. (1997). *The Early Irish Verb* (Rev. 2nd ed. with index verborum.). An Sagart.

Ó Corráin, D. (2017). *Clavis Litterarum Hibernensium: Medieval Irish Books & Texts (c. 400 – c. 1600)* (Vol. *1–3*). Brepols.

Qiu, F., & Stifter, D. (2020). Chronologicon Hibernicum: Frámaíocht dhóchúlaíoch chun dátú a dhéanamh ar fhorbairtí i dteanga na Sean-Ghaeilge [Chronologicon Hibernicum: A probabilistic framework for the dating of Old Irish language developments]. In E. Ó Raghallaigh (Ed.), *Téamaí agus Tionscadail Taighde* (pp. 39–59). An Sagart.

Qiu, F., Stifter, D., Bauer, B., Lash, E., & Tianbo, J. (2018). Chronologicon Hibernicum: A probabilistic chronological framework for dating Early Irish language developments and literature. In M. Ioannides et al. (Eds.), *Digital Heritage: Progress in Cultural Heritage: Documentation, Preservation, and Protection* (pp. 731–740). Springer. https://doi.org/10.1007/978-3-030-01762-0_65

R Core Team (2020). *R: A Language and Environment for Statistical Computing* (Version 4.0.0) [Computer Software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rögnvaldsson, E., & Helgadóttir, S. (2011). Morphosyntactic tagging of Old Icelandic texts and its use in studying syntactic variation and change. In C. Sporleder, A. Bosch, & K. Zervanou (Eds.), *Language Technology for Cultural Heritage* (pp. 63–76). Springer. https://doi.org/10.1007/978-3-642-20227-8_4

Sagart, L., Jacques, G., Lai, Y., Ryder, R. J., Thouzeau, V., Greenhill, S. J., & List, J. (2019). Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proceedings of the National Academy of Sciences of the USA 116*(21), 10317–10322. https://doi.org/10.1073/pnas.1817972116

Schneider, G. (2008). *Hybrid Long-Distance Functional Dependency Parsing* [Doctoral dissertation, University of Zurich]. https://www.cl.uzh.ch/dam/jcr:ffffffff-c155-5f61-0000-00004dc66d11/schneider_diss.pdf

Schreier, D. (2005). #CCV- > #CV-: Corpus-based evidence of historical change in English phonotactics. *International Journal of English Studies*, *5*(1), 77–99.

Schumacher, S. (2004). *Die keltischen Primärverben: Ein vergleichendes, etymologisches und morpho-logisches Lexikon* [The Celtic Primary Verbs: A Comparative, Etymological and Morphological Dictionary]. Innsbruck.

Stifter, D. (2009). Early Irish. In M. Ball & N. Müller (Eds.), *The Celtic Languages* (2nd ed., pp. 55–116). Routledge.

Stifter, D., Barrett, S., Bauer, B., Ganly, E., Griffith, A., Ji, T., Lash, E., Nguyen, T. H., Osarobo, G., Qiu, F., & White, N. (2021–). *Corpus Palaeohibernicum*. https://chronhib.maynoothuniversity.ie/chronhibWebsite/

Stokes, W., & Strachan, J. (Eds.). (1901–1910). *Thesaurus Palaeohibernicus: A Collection of Old Irish Glosses, Scholia, Prose and Verse*. Dublin Institute for Advanced Studies.

Su, Y.-S., & Yajima, M. (2020). *R2jags: Using R to Run 'JAGS'* (Version 0.6–1). https://CRAN.R-project.org/package=R2jags

Rama, T., & Wichmann, S. (2020). A test of generalized Bayesian dating: A new linguistic dating method. *PLOS ONE 15*(8): e0236522. https://doi.org/10.1371/journal.pone.0236522

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

Toner, G., & Han, X. (2019). *Language and Chronology: Text Dating by Machine Learning*. Brill. https://doi.org/10.1163/9789004410046

Uhlich, J. (2018). Review article of: P. Ó Riain (ed.), *The Poems of Blathmac Son of Cú Brettan: Reassessments*. Irish Texts Society, 2015. *Cambrian Medieval Celtic Studies*, *75*, 53–77.

Address for correspondence

David Stifter
Department of Early Irish
Maynooth University
Maynooth, Co. Kildare
Ireland

david.stifter@mu.ie
https://orcid.org/0000-0001-5634-9912

Co-author information

Fangzhe Qiu
School of Irish, Celtic Studies and Folklore
University College Dublin

fangzhe.qiu@ucd.ie
https://orcid.org/0000-0002-7167-9001

Marco A. Aquino-López
Centro de Investigación en Matemáticas
Universidad de Guanajuato

aquino@cimat.mx
https://orcid.org/0000-0002-5076-7205

Bernhard Bauer
Centre for Information Modelling – Austrian
Centre for Digital Humanities
Karl-Franzens-Universität Graz

bernhard.bauer@uni-graz.at
https://orcid.org/0000-0003-2881-0972

Elliott Lash
Sprachwissenschaftliches Seminar
Georg-August-Universität Göttingen

elliottjamesfrick.lash@uni-goettingen.de
https://orcid.org/0000-0001-7004-471X

Nora White
Department of Early Irish
Maynooth University

nora.white@mu.ie
https://orcid.org/0000-0001-7957-651X

Publication history

Date received: 14 April 2021
Date accepted: 4 February 2022
Published online: 20 September 2022