







# *Chronologicon Hibernicum*: A Probabilistic Chronological Framework for Dating Early Irish Language Developments and Literature

Fangzhe Qiu<sup>1</sup>(✉) , David Stifter<sup>1</sup> , Bernhard Bauer<sup>1</sup> ,  
Elliott Lash<sup>1</sup>, and Tianbo Ji<sup>2</sup> 

<sup>1</sup> Maynooth University, Maynooth, Co. Kildare, Ireland  
{fangzhe.qiu, david.stifter, bernhard.bauer,  
elliott.lash}@mu.ie

<sup>2</sup> Dublin City University, Dublin, Ireland  
tianbo.ji2@mail.dcu.ie

**Abstract.** This paper introduces the ongoing ERC-funded project *Chronologicon Hibernicum*, which studies the diachronic developments of the Irish language between c. 550–950, and aims at refining the absolute chronology of these developments. It presents firstly the project organization, its subject matter and objective, then gives an overview of the potentials and challenges in studying the Early Irish language. The project combines historical linguistic analysis, corpus linguistic methods and Bayesian statistic tools. Finally the paper explains the impact of this project in preserving the Irish cultural heritage and the lessons learned in the first three years.

**Keywords:** Chronologicon hibernicum · Linguistic dating  
Irish cultural heritage

## 1 Introduction to the Project

### 1.1 Basic Facts

The research project ‘Chronologicon Hibernicum – A Probabilistic Chronological Framework for Dating Early Irish Language Developments and Literature’ has received funding through a Consolidator Grant of the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 647351). It started in September 2015 and will continue to the end of August 2020. It is hosted in the Department of Early Irish, Maynooth University, Ireland. The project team currently consists of the Principal Investigator (Prof. David Stifter), three full-time postdoctoral researchers (Dr. Bernhard Bauer, Dr. Elliott Lash and Dr. Fangzhe Qiu), two PhD students (Romanas Bulatovas and Lars Nooij) and two research assistants (Ellen Ganly and Tianbo Ji) and will soon welcome an extra staff member.

## 1.2 Subject Matter

The two words *Chronologicon Hibernicum* contain the main aspects of the project in a nutshell. *Chronologicon* is a Greek adjective with the meaning ‘pertaining to chronology’. So it is time and the impact that time leaves on language which lies at the heart of this project. *Hibernicum* is the Latin adjective for ‘Irish’. The subject matter of *Chronologicon Hibernicum*, or ChronHib in short, is the **diachronic development of the Early Irish language**. Languages change over time, and in all linguistic domains, especially phonology, morphology and syntax, developments of the Irish language are clearly visible in the form of variations through time [1]. In absolute dates, the period studied in the project falls between c. 550 and c. 950 CE, covering what are traditionally termed the Early Old Irish, Old Irish and Early Middle Irish periods.

## 1.3 Objectives

The project’s central objective is **linguistic dating**, i.e. to link the changes in linguistic forms to certain periods of time. On the one side one asks, when was an older form A replaced by the newer form B; on the other, at a given point of time T, what is the probability that one finds B instead of A in same context? If these two questions can be answered, then one can predict the age of a text by examining its linguistic profiles.

Traditionally, linguistic dating is done by philological and linguistic analysis of manually curated data. ChronHib aims at revolutionizing the methods used for linguistic dating of Early Irish, by contributing to a chronologically more precise description of the variations in the above-mentioned linguistic domains, and by employing corpus linguistic and advanced statistical methods. It also endeavors to improve, by means of digital humanities techniques, on the availability and reliability of the material basis relevant to the chronology of linguistic developments and of the literature of Early Medieval Ireland.

## 2 The Early Irish Language: Potentials and Challenges

Since the diachronic development of Early Irish is the subject matter of ChronHib, it is pertinent to give a brief introduction to the Early Irish language and its textual culture.

The Irish language belongs to the Celtic branch of the Indo-European language family, and its closest relatives are Welsh, Breton, Cornish and ancient languages such as Gaulish and Celtiberian. Early Irish refers to the phases of the Irish language before c. 1200 CE, including that of Primitive Irish (before c. 550), Old Irish (c. 550–900) and Middle Irish (c. 900–1200). Middle Irish is followed by Early Modern Irish (c. 1200–1700 CE) [2]. Early Irish was mainly spoken in the island of Ireland, but was also used in the Irish colonies in Britain and the Isle of Man. It is the ancestor of modern Irish, Scottish Gaelic and Manx.

Excluding formulaic inscriptions and proper names recorded by Classical authors, the earliest evidence of written Early Irish dates to the 7<sup>th</sup> century, in Latin script brought to Ireland by Christianity. From the 7<sup>th</sup> to the 10<sup>th</sup> century, the written literary tradition was vast in sheer number of texts and variegated in extent and genres. The

number of extant texts from this period can be estimated to between eight and nine hundred prose texts of quite diverse length, and this does not even include the extremely rich poetic tradition. The texts include narrative sagas, historical texts (annals, genealogies), pseudo-historical tales, religious writings (homilies, saints' lives, martyrologies), poetry, as well as an extensive learned tradition of law, medicine, grammar and computistics, produced both in Ireland and in the multicultural Irish monasteries on the Continent.

These texts constitute a rich and unique cultural heritage not only of Ireland, but of Western Europe at large. They provide a detailed picture of the social, political and intellectual lives in early medieval Ireland, from the power struggles between kingdoms and diverse theological theories, down to the stories behind local place-names and regulations for bee-keeping. They give an indispensable account of a time when in continental Europe very little of other vernaculars had been rendered into letters, and contemporary records are sparse and obscure. More importantly, they testify to the thriving intellectual activities in Ireland and other parts of Europe, and the close connection between them. Since the majority of these texts are written in Early Irish, the study of Early Irish is quintessential for realizing the full potential of this cultural heritage.

However, Early Irish texts pose many challenges to modern scholars. The first one is the inherent characteristics of Early Irish as a morphologically complex language. Although the inventory of Early Irish inflection is similar to that of Latin or other ancient Indo-European languages, prehistoric phonological changes have rendered the synchronic morphological rules opaque and irregular. For non-specialists, it would be difficult, for instance, to recognize the lexeme *orgaid* '(s)he kills' in *iurtair* 'they will be killed' even with the help of grammars and dictionaries. There are still many gaps and obscurities, even for linguists, in the knowledge of Early Irish, such as the phonological rules that govern the change from *aue* 'grandson' to *ó*, or the reason of occasional omission of the relative particle *-(s)a* in prepositional relative clauses.

The second difficulty lies in the fact that, despite their richness in extent and content, little is known about the historic contexts in which the texts were produced. Almost all of the literature, especially of the early period, has been transmitted anonymously, and for most texts the time and circumstances of composition are unknown. The problem is compounded by transmission. Texts are materialized in the form of manuscripts, yet the absolute majority of medieval Irish texts today exist only in manuscripts made in or after the 12<sup>th</sup> century, and these texts occasionally underwent substantial revision or modernization on their way through history. Episodes of the famous heroic saga *Táin Bó Cúailnge* ('the Cattle-Raid of Cooley'), may first have been written down in the 8<sup>th</sup> century and were then joined with other episodes in the 9<sup>th</sup> century. What we have, however, is a copy from the early 12<sup>th</sup> century that may well have been partly adapted to the orthography and to the pronunciation of that time. Then we have another completely reworked version, in a very bombastic style, written in the latter part of the 12<sup>th</sup> century. In a mix like this it is very challenging to disentangle the complex sequence of chronological layers, although most texts are homogenous enough even if they only exist in later copies.

The two challenges, however, can be effectively tackled by fine-tuned linguistic analysis and sufficient linguistic dating. Mining and comparing linguistic data from

different periods provide quantifiable measures to the changes in the language, thereby revealing previously unknown or unclear grammatical rules. All of these contribute to our deeper understanding of the nature and development of Early Irish, and help us interpret Early Irish texts more accurately. Meanwhile, close dating of the language makes it possible to identify the period in which a text was written, therefore linking a text to a specific historical setting. Texts can then be put into a precise chronological order, which enables us to trace their transmission and evolution that can reveal a lot about the material and intellectual culture of the time. Linguistic dating elucidates the crucial parameter of time in historical research and therefore is paramount for a better appreciation of the cultural heritage.

### 3 Methodologies

#### 3.1 Data Collection

Since the Early Irish language is preserved as texts written in medieval manuscripts, these texts provide the data for linguistic dating in ChronHib. They are further categorized as below:

**Texts from Contemporary Manuscripts.** These are texts written in manuscripts produced before 950. Given the above-mentioned active scribal intervention in the transmission of earlier texts, only texts in manuscripts written before 950 can be trusted as accurately reflecting the linguistic characteristics of Irish in the period under examination. Around 80 texts are known to belong to this category, mostly in manuscripts now kept on the continent [3]. However, the date of the production of the manuscript is not always precisely known, and the relationship between the text and the manuscript sometimes remains obscure. A text on the life of St. Columba was composed by Adomnán of Iona between 688 and 692, and a copy was made by Dorbéine before he died in 713, which is now kept in Schaffhausen, Switzerland [4]. In this case both the manuscript and the text can be closely dated. Milan, Biblioteca Ambrosiana, C 301 inf. contains a copy of the commentary on the Psalms that has been heavily glossed in Irish. This manuscript can be dated to the first quarter of the 9<sup>th</sup> century, but the glosses seem to have been copied from an earlier source [3].

**Texts That Are Non-linguistically Dated but from Later Manuscripts.** These are texts that can be dated on non-linguistic grounds to specific periods but are only found in manuscripts produced after 950. A salient example is the ‘Law of the Innocents’ by Adomnán, the promulgation of which in June, 697 was recorded in the annals [5]. Yet the only copies that survive are from the 15<sup>th</sup> and the 17<sup>th</sup> centuries respectively. The poems by Blathmac son of Cú Bretan are found in a single 17<sup>th</sup>-century manuscript, although the annals report his father’s death in 740 [5]. In these cases, scribal modernization has often affected the orthography and to some extent other linguistic features as well, though the rhymes in verses frequently help us restore the original forms. As a result, these texts are of less evidential value to the linguistic profile of Irish between c.550–950 than the first category, but are still invaluable data.

**Texts That Cannot Be Dated Non-linguistically.** One can generally say that these texts belong to the Early Irish period, even attempt to assign them to specific centuries, judging from their linguistic appearances, but these dates are often impressionistic and too broad for linguistic dating purposes. Nor are these texts found in manuscripts before 950. Consequently, these texts are not used as data for creating diachronic linguistic profiles.

### 3.2 Pre-processing the Data

Texts from the first and second categories listed in 3.1 constitute the corpus of data used in ChronHib. They are subject to the following pre-processing procedures. Most of the procedures are so far done manually or semi-automatically with search and replace commands, but we are developing fully automatic taggers.

**Digitalization.** Several digitalized corpora of Early Irish texts have already been published, which can be directly incorporated into the ChronHib database. These include corpora that are already linguistically parsed, such as the *Milan Glosses Database* [6], the *Priscian Glosses Database* [7], and the *Parsed Old and Middle Irish Corpus* [8], as well as a number of text repositories, such as the *Corpus of Electronic Texts* [9] and *Thesaurus Linguae Hibernicae* [10]. Other target texts have either been edited in the two-volume *Thesaurus Palaeohibernicus* [11], or have been published in individual critical editions, such as the Patrician texts in the Book of Armagh [12], or the *Vita Sancti Columbae* [4]. These edited texts are OCR-ed into digital format, and are proofread against the manuscript images.

**Tokenization.** Texts are broken down into sentences or glosses, and further into individual tokens consisting of minimally analyzable lexical units called ‘morphs’. For example, the verbal complex *arnacha-toirsitis* ‘so that they might not take her’ [6] (48d27) is tokenized into *ar* ‘so that’, *nach* ‘that not’, *a* ‘her’, *to* (preverb), *r* (augment) and *toirsitis* ‘they might take’ respectively. To date the ChronHib corpus consists of 111,272 tagged tokens from 69 dated texts, and is still expanding.

**Lemmatization, POS- and Morphological Tagging.** Each token is assigned a lemma (the citation form of a lexeme), and given tags on its part-of-speech (POS) and morphological information according to a unified tagset, as exemplified in Table 1. Other information, such as etymology, mutation or onomastic compounds, is also annotated when applicable.

**Variation-Tagging.** A table has been created that lists 326 linguistic variations that we have currently identified to have occurred in the Irish language during the period c. 550–950. These include phonological, orthographical, morphological, syntactical and lexical variations. Each variation is given an ID (e.g. PH030) and a description stating the possible values of the variable (e.g. ‘pretonic /e/becomes /a/’). For each variable, the linguistic condition is defined (e.g. in the pretonic position in PH030), the values are usually binary (e.g. /e/vs. /a/in PH030), and sometimes the chronological order of the values is known (e.g. /e/is earlier than /a/in PH030).

Every token in the corpus is then tagged as to: (1) which variation could have possibly happened in the linguistic condition provided by the token, and (2) which

**Table 1.** Examples of lemmatization and POS-tagging in the *ChronHib* corpus.

Morph	Lemma	POS	Classification	Gender	Meaning	Morphological analysis
<i>ar</i>	<i>ara 1</i>	conjunction			so that, in order that	
<i>nach</i>	<i>nád 1</i>	particle	relative		that not	
<i>a</i>	<i>3sg.fem. inf.pron.</i>	pronoun	infixd		her	Class C
<i>to</i>	<i>do-</i>	particle	preverb			
<i>r</i>	<i>ro 1</i>	particle	augment		perfective or potential aspect	
<i>-toirsitis</i>	<i>do-fich</i>	verb	S1		to take, to attack	aug.3pl.past.subj.

value of the variable does the token show. One of the binary values of the variable, normally the earlier one if chronological order is known, is tagged **No**, while the other value **Yes**. Unclear instances are tagged **Maybe**. Table 2 offers some examples:

**Table 2.** Examples of variation-tagging in the *ChronHib* corpus.

Morph	Var.ID	Var.Description	Value
<i>Achid</i>	OR005	use <ai> instead of earlier <i> to represent the schwa in the unstressed syllable CəC'	<b>No</b>
<i>das</i>	MO072	use of new infixd pronoun forms instead of old ones	<b>Yes</b>
<i>Feradach</i>	PH029	posttonic, non-final short vowels are reduced to schwa	<b>Maybe</b>

### 3.3 Data Analysis

**Synchronic Linguistic Profiles.** By means of the close tagging described in Sect. 3.2, qualitative linguistic information can be transformed into a quantitative one. We can produce a numerical account of the linguistic variations of a text that has been dated by non-linguistic criteria. Since such a text represents the linguistic reality of a certain period, we can use the account of variations as a synchronic linguistic profile of that period. For instance, Table 3 shows the number of tags for a few phonological variations in the Schaffhausen copy of *Vita Sancti Columbae* (688x713), which tells us that at the end of the 7<sup>th</sup> century, while some changes (e.g. PH010 and PH025) have not yet started to happen in Irish (given that the innovative form tagged by **Yes** does not occur at all, the percentage of **Yes** being 0%), other changes have already begun (e.g. PH008) or have reached completion (e.g. PH028) (given that the older form tagged by **No** does not occur at all, the percentage of **Yes** being 100%).

When texts from different periods are tagged for linguistic variations, we have individual synchronic profiles of the Irish language from these periods. It has to be

**Table 3.** Examples of variations in *Vita Sancti Columbae*

ID	n (tokens)	No	Yes	Maybe	Yes percentage in all tokens
PH006	16	15	1	0	6.25%
PH008	212	147	41	24	19.34%
PH010	13	13	0	0	0.00%
PH013	145	136	1	8	0.69%
PH015	25	24	1	0	4.00%
PH025	30	30	0	0	0.00%
PH028	17	0	17	0	100.00%

remembered that these profiles are not continuous or exhaustive: they constitute random samples of a constantly changing language during c. 550–950.

**Statistical Analysis.** Because language changes are by nature probabilistic and cumulative rather than categorical and abrupt, the linguistic profiles are expressed by frequencies, both of the values of variations, and of the appearance or absence of certain forms and structures. Moreover, the periods that can be profiled are neither continuous nor evenly distributed. Therefore statistical methods must be employed, especially Bayesian statistics, which allows to make statements about prior knowledge in the light of newer information, i.e. ‘degrees of belief’ about propositions whose truth or falsity is uncertain can enter the equation [13, 14].

Synchronic profiles are combined to form a diachronic linguistic profile that consists of three major variables: date, variation and the number of tags. The statistical analysis will serialize the numbers of tags per variation according to the dates, and run multi-variable regression to create an absolute chronology of linguistic changes in Irish, which will inform us of the probabilities of certain linguistic features at any given temporal point within the investigated period. The detailed statistical methods are to be developed later in the project.

**Testing of the Absolute Chronology.** A small portion (about 10%) of tokens from dated texts will not join the statistical analysis. They are reserved as control data for testing the accuracy of the absolute chronology. These will be profiled separately, and their profiles will be mapped onto the absolute chronology to calculate their possible dates. If the predicted date matches the actual date (margin of error allowed), then the absolute chronology is valid; if not, the new data from the control group will be used to improve the calculation.

### 3.4 Application to Undated Texts

If proven sufficiently accurate, the absolute chronology framework can then be used to predict the date of a hitherto undated Irish text. An undated text undergoes the pre-processing as specified in Sect. 3.2, and a synchronic linguistic profile is created for it. The profile is then subject to multi-variable statistic tests to calculate the probability of its date by comparing it to the profiles of texts of known dates and to the absolute chronology. The result will be the confidence interval (or credible interval in Bayesian

statistics) of the date of the text, interpreted as a range of years with corresponding probability. We will try to achieve a balance between the precision of date and the confidence level. The text can thus be quantitatively linguistically dated. Again, the actual statistical methods for this step remain to be developed at a later stage of the project.

## 4 Technology

The digitalization of data employs OCR scanning and translation of TEI and HTML files into .csv and .xlsx formats. We use Python scripts to tokenize Early Irish texts, and the tokens are then imported into a database developed on the FileMaker™ software for lemmatization and tagging.

Data processed on FileMaker™ are then exported, by the help of Python scripts, to a server-based database built upon MySQL. MySQL is the most popular free open-source relational database management system. Users can manage MySQL with MySQL Workbench – a unified visual tool for database developers which provides data modeling, SQL development, and comprehensive administration tools for server configuration, user administration, backup, and much more.

The project website uses HTML5 and Flask. On the frontend, we use a free Bootstrap website template as our index page, in which HTML, CSS and JavaScript (including pure JavaScript, jQuery and Ajax) are introduced. The backend is Flask, a Python website micro-framework based on Werkzeug and Jinja 2.

For the server side, the website and database are deployed on Maynooth University's Apache Server in which we can automatically back up our data and rollback if error occurs.

## 5 Impact and Expected Outcomes

The ChronHib database will soon be online for open access [15]. It is by far the largest linguistically annotated digital corpus of Early Irish texts. It serves as an electronic archive, which can be freely browsed and searched by anyone interested in early medieval history, literature, scribal practice or any other related fields. As an intensively annotated linguistic corpus, it also appeals not only to researchers of Early Irish, but also to linguists further afield as data for comparative or general linguistic studies.

This annotated corpus will also be the basis of an automatic tagger program for Early Irish texts. Trained by existing data and equipped with machine-learning techniques, this tagger will be able to annotate Early Irish texts (morphological, POS or syntactic) with a high accuracy.

In the process of building the corpus, we have established various standards and ontologies in collecting, annotating and analyzing data of Early Irish. The tagset developed by the project, for instance, is at present the most efficient and comprehensive for Early Irish. Our method of variation tagging is innovative, not only in the discipline of Early Irish, but also in diachronic linguistics at large. These formal expressions and methodologies are valuable assets and will benefit future researches.



The absolute chronology of linguistic developments in Early Irish and the statistical models will be the most important outcome of this project. Many new insights will be gained into the Early Irish language, which is a crucial component of the Irish culture heritage. The language is also the key to understanding the intellectual history and the textual culture in the Irish cultural sphere. ChronHib will create an authoritative reference point for linguistic dating, which will assign trustworthy dates to texts of medieval Ireland, thereby unravelling the complex intertextuality of Irish literature. Because Early Irish is beset with all the typical problems of Natural Language Processing on a historical language, such as small size of corpus, unstandardized spelling, morphological complexity and imbalance of registers, the statistical models developed for Early Irish will greatly advance the toolkit for processing other historical languages as well. These outcomes will be presented in the form of scholarly articles and books.

## 6 Lessons Learned

The lack of a sufficiently precise standard for the linguistic analysis of Old Irish has been a major delaying factor. There is no uniform system of tagging in the standard dictionary of Old Irish, nor in existing printed and digital editions. As a consequence, data cannot simply be imported from pre-existing collections. Likewise, the pre-existing databases based on Filemaker™, on which we built our initial corpus, tagged Early Irish texts slightly differently as they had been designed specifically for individual texts. We only realized after a while that the structures and tagsets of these databases do not suit the diversity and tagging needs of our corpus. In addition to this, since we used individually installed Filemaker™, the annotation practices varied from one member to another. It has taken us a very long time afterwards to harmonize all previous works into a uniformed format. Looking back, we should have established our standard structure and tagset and built the server-based database at the very beginning to avoid wasting time in harmonizing them at a later stage.

A related problem consists in the fact that the textual editions that our corpus is based on turned out to be less reliable philologically than we had assumed initially. The alternatives to cope with this problem are, either to simply accept errors into the analysis of our corpora, or to use the opportunity of corpus-building to improve the texts philologically, which, however, slows down the tagging process.

**Acknowledgements.** This paper is written as part of the research project *Chronologicon Hibernicum*, which has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 647351).

## References

1. Stifter, D.: Towards the linguistic dating of early Irish law texts. In: Ahlqvist, A., O'Neill, P. (eds.) *Medieval Irish Law: Text and Context*, pp. 163–208. The University of Sydney, Sydney (2013)
2. McCone, K.: *Towards a relative chronology of ancient and medieval Celtic sound changes*. Maynooth University, Maynooth (1996)

3. Bronner, D.: Verzeichnis altirischer Quellen. Philipps Universität Marburg, Marburg (2013)
4. Anderson, A.O., Anderson, M.O. (eds.): Adomnan's Life of Columba. Thomas Nelson & Sons, London (1961)
5. Mac Airt, S., Mac Niocaill, G. (eds.): The Annals of Ulster (to A.D. 1131). Dublin Institute for Advanced Studies, Dublin (1983)
6. Griffith, A., Stifter, D.: Dictionary of the Old Irish glosses in the Milan MS Ambr. C301 inf. [http://www.univie.ac.at/indogermanistik/milan\\_glosses.htm](http://www.univie.ac.at/indogermanistik/milan_glosses.htm). Accessed 13 Aug 2018
7. Bauer, B.: The online database of the Old Irish Priscian glosses. <http://www.univie.ac.at/indogermanistik/priscian/>. Accessed 13 Aug 2018
8. Lash, E.: The Parsed Old and Middle-Irish Corpus. <https://www.dias.ie/celt/celt-publications-2/celt-the-parsed-old-and-middle-irish-corpus-pomic/>. Accessed 13 Aug 2018
9. Corpus of Electronic Texts. <https://celt.ucc.ie//index.html>. Accessed 13 Aug 2018
10. Thesaurus Linguae Hibernicae. <http://www.ucd.ie/tlh/>. Accessed 13 Aug 2018
11. Stokes, W., Strachan, J. (eds.): Thesaurus Palaeohibernicus. Dublin Institute for Advanced Studies, Dublin (1987)
12. Bieler, L.: The Patrician Texts in the Book of Armagh. Dublin Institute for Advanced Studies, Dublin (1979)
13. Bayes, T.: An essay towards solving a problem in the doctrine of chances. *Philos. Trans. Roy. Soc. London* **53**, 370–418 (1763)
14. Malakoff, D.: Bayes offers a 'new' way to make sense of numbers. *Science* **286**, 1460–1464 (1999)
15. ChronHib. <http://chronhib.maynoothuniversity.ie>. Accessed 13 Aug 2018