Elliott Lash, Fangzhe Qiu, David Stifter (Eds.)
Morphosyntactic Variation in Medieval Celtic Languages

Trends in Linguistics
Studies and Monographs

Volume 346

Morphosyntactic Variation in Medieval Celtic Languages

Corpus-Based Approaches

Edited by
Elliott Lash, Fangzhe Qiu, David Stifter

DE GRUYTER
MOUTON

This book was written as part of the project *Chronologicon Hibernicum*. 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). The editors of this volume also thank the Maynooth University Publications Fund for providing financial support for the publication of this volume in March, 2020, and the National University of Ireland Publications Scheme for providing financial support in July, 2020.

ISBN 978-3-11-068066-9
e-ISBN (PDF) 978-3-11-068074-4
e-ISBN (EPUB) 978-3-11-068079-9
DOI https://doi.org/10.1515/9783110680744

Library of Congress Control Number: 2020940693

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Typesetting: Integra Software Services Pvt. Ltd.
Printing and binding: CPI books GmbH, Leck

www.degruyter.com

Contents

List of contributors

The papers in this book arose from lectures given by the following contributors, hosted at the Department of Early Irish, Maynooth University by the project *Chronologicon Hibernicum* (Maynooth University, ERC Consolidator Grant 2015, H2020 #647351).

Liam Breatnach, MRIA is a Senior Professor at the School of Celtic Studies, Dublin Institute for Advanced Studies, and co-editor of the journal *Ériu*, published by the Royal Irish Academy. His main research interests are Old Irish, Middle Irish and the historical development of Irish, law texts, and poets, poetry and metrics. A recent publication on Early Irish law is *Córus Bésgnai* (Breatnach 2017a). A recent publication on poetry and metrics is about the *Trefocal* tract (Breatnach 2017b). A recent publication on language is "Lebor na hUidre: Some linguistic aspects" (Breatnach 2015).

Benjamin Bruch teaches literature, history, and linguistics at the Pacific Buddhist Academy (Honolulu, Hawai'i). His research interests include metrics and versification, the historical phonology and syntax of the Celtic languages, and the preservation and revitalisation of endangered languages. Bruch received a Ph.D. in Celtic Languages and Literatures at Harvard University with a dissertation titled *Cornish verse forms and the evolution of Cornish prosody, c. 1350–1611* (Bruch 2005). He has published on medieval Cornish literature in verse: "Medieval Cornish versification: An overview" (Bruch 2009). He has also worked on Cornish historical phonology: "Nucleus length and vocalic alternation in Cornish diphthongs" (Bock and Bruch 2010) and "New perspectives on vocalic alternation in Cornish" (Bock and Bruch 2012). Bruch is also a co-author of the new standard orthography for Revived Cornish: *An outline of the standard written form of Cornish* (Bock and Bruch 2008), officially adopted by the Cornish Language Partnership.

Carlos García-Castillero teaches various subjects related to Indo-European linguistics and historical and comparative linguistics at the Faculty of Arts of the University of the Basque Country. He studied Classical Philology at the University of the Basque Country and wrote his Ph.D. thesis on Indo-European linguistics (*La formación del tema de presente primario osco-umbro*; García-Castillero 1999) under the supervision of Prof. Jürgen Untermann and Prof. Joaquín Gorrochategui. His main fields of research are comparative Indo-European linguistics (with several papers on pronominal and verbal morphosyntax), Old Irish morphosyntax (illocutionary force and clause types, templatic character of the verbal complex), and pragmatics, as well as diachronic linguistics. His publications can be found in a wide range of journals, such as *Ériu*, *Zeitschrift für celtische Philologie*, *Historische Sprachforschung*, *Indogermanische Forschungen*, *Journal of Historical Pragmatics*, and *Diachronica*.

Joseph F. Eska is Professor of Linguistics at the Department of English at Virginia Polytechnic Institute and State University. He has worked on all aspects of the linguistic history of the Celtic languages, in particular on the ancient Celtic languages of Continental Europe. He is the author (with Don Ringe) of *Historical linguistics: Toward a twenty-first century reintegration* (Ringe and Eska 2012) and over 80 articles and book chapters on diachronic Celtic linguistics. He is currently completing a monograph on the syntax of the Continental Celtic languages and is investigating the architecture of the left periphery in the Insular Celtic languages within a

https://doi.org/10.1515/9783110680744-203

Cartographic framework. He is the editor of the *North American Journal of Celtic Studies* and co-editor of *Indo-European Linguistics*.

Theodorus Fransen was awarded his Ph.D. from Trinity College Dublin in 2020 for a thesis entitled *Past, present and future: Computational approaches to mapping Old and Modern Irish cognate verb forms*. He is currently a postdoctoral researcher on the Cardamom project (Comparative deep models for minority and historical languages), led by John P. McCrae, at the Data Science Institute, National University of Ireland, Galway. While his focus so far has been on computational morphology for Old Irish verbs, he has a growing interest in the broader area of Natural Language Processing and its applications in fields such as lexicography and Digital Humanities. He hopes to explore new, digital avenues to facilitate systematic investigation of linguistic developments between Old and Modern Irish in a future research project. His forthcoming publications include "Automatic morphological parsing of Old Irish verbs using finite-state transducers" (to appear in *Leeds Working Papers in Linguistics and Phonetics*).

Aaron Griffith is Assistant Professor at the Department of Celtic Languages and Culture of Utrecht University. After having finished a Ph.D. at the University of Chicago on various problems in Insular Celtic historical phonology and morphology, he took up a postdoctoral position at the University of Vienna, where he created a digital database and new edition of the Milan Glosses (Griffith and Stifter 2013). The glosses remain a research interest of his, as do the pronominal systems, syntax, and typological profiles of the Insular Celtic languages.

Britta Irslinger is a researcher in the project *Deutsche Wortfeldetymologie in europäischen Kontext* at the Saxon Academy of Sciences and Humanities in Leipzig. She is an Indo-Europeanist and Celticist. Her dissertation on abstract nouns with dental suffixes in Old Irish (*Abstrakta mit Dentalsuffixen im Altirischen*) appeared in 2002. Her major publications include *Nomina im indogermanischen Lexikon* (Wodtko, Irslinger and Schneider 2008) and articles on various topics in historical and comparative linguistics, such as "The gender of abstract noun suffixes in the Brittonic languages" (Irslinger 2014a) and "More tales of two copulas" (Irslinger 2019). Further topics are pre-modern and modern concepts of Proto-Indo-European and Celtic Culture and the history of linguistics, cf. "Medb 'the intoxicating one'?" (Irslinger 2017a) and "Geographies of identity" (Irslinger 2017b). Her contribution to this book arises from a previous project, "Detransitivity in the Brittonic languages: reflexivity, reciprocity and Middle voice constructions".

Marius L. Jøhndal obtained his Ph.D. in 2012 from the University of Cambridge and was previously employed at the University of Oslo, where he worked on the PROIEL and Syntacticus treebanks of Indo-European languages. His research has focused on Latin syntax, in particular non-finiteness and reflexivity, and on computational methods for historical linguistics. He currently works for Google.

Elliott Lash is a postdoctoral researcher on the ERC-funded *Chronologicon Hibernicum* project at Maynooth University. He obtained his Ph.D. from the University of Cambridge in 2011 for a thesis entitled *A synchronic and diachronic analysis of Old Irish copular clauses*. Afterwards, he was an O'Donovan Scholar at the Dublin Institute for Advanced Studies from 2011 to 2014, and a ZIF-Marie Curie Fellow at the University of Konstanz from 2014 to 2016. His research

interests are syntax and language change, with a special focus on the history of the Irish language. He is currently writing an introduction to Old Irish syntax. Major journal articles published by him include "Coordinate subjects, expletives, and the EPP in early Irish" (Lash and Griffith 2018), "A quantitative analysis of e/i variation in Old Irish *eter* and *ceta*" (Lash 2017a), "Evaluating directionality in the internal reconstruction of pre-Old Irish copular clauses" (Lash 2017b), and "Subject positions in Early Irish" (Lash 2014b).

Marieke Meelen is a postdoctoral researcher and affiliated lecturer at the University of Cambridge and Fellow-Commoner at Trinity Hall. After completing her Ph.D. at Leiden University on word order change in the history of Welsh in June 2016, she moved to the UK for a postdoc in the ReCoS project led by Prof. Ian Roberts, working on comparative Tibeto-Burman syntax. After the completion of the project, she was awarded a British Academy postdoctoral fellowship to work on her own project "The emergence of V2 word order", combining information structure and historical syntax with NLP techniques and building the Parsed Historical Corpus of the Welsh Language (PARSHCWL) and the Annotated Corpus of Classical Tibetan (ACTib).

Fangzhe Qiu received his Ph.D. degree in Early and Medieval Irish from University College Cork in 2015. He was an O'Donovan Scholar at the Dublin Institute for Advanced Studies from 2014 to 2015 and then a postdoctoral researcher in the project *Chronologicon Hibernicum* in the Department of Early Irish, Maynooth University, Ireland from 2015 to 2019. He is currently a lecturer of Celtic Studies at University College Dublin. He has published widely on early Irish law, Old Irish language and medieval Irish manuscripts. His current research interests include medieval Irish annals, Celtic languages and quantitative historical linguistics. Some of his recent publications are: "Old Irish *aue* 'descendant' and its descendants" (Qiu 2019), "The first judgment in Ireland" (Qiu 2018), and "The Ulster Cycle in the law tracts" (Qiu 2017).

Elisa Roma is Associate Professor of Linguistics at the University of Pavia, where she earned her Ph.D. in Linguistics in 1998. The results of her Ph.D. research were published in a monograph (*Da dove viene e dove va la morfologia*; Roma 2000a) and an article ("How subject pronouns spread in Irish"; Roma 2000b). Her main research interests cover typologically oriented historical and comparative linguistics, Celtic philology and in particular Gaelic morphosyntax. She is currently involved in the Italian National Project "Transitivity and argument structure in flux" (Universities of Naples and Pavia). Her strong commitment to multilingualism has led her to translate scholarly works in Irish (*L'irlandese antico e la sua preistoria* and *Il medioirlandese/Middle Irish*; respectively McCone and Roma 2005; Breatnach and Roma 2013). Recent publications include: *Linguistic and Philological Studies in Early Irish* (Roma and Stifter 2014), "Nasalization after inflected nominals in the Old Irish glosses: A reassessment" (Roma 2018a), "Old Irish pronominal objects and their use in verbal pro-forms" (Roma 2018b), "On the origin of the absolute vs. conjunct opposition in Insular Celtic" (Budassi and Roma 2018).

David Stifter is Professor of Old and Middle Irish at Maynooth University. He is founder and editor of the interdisciplinary Celtic Studies journal *Keltische Forschungen* (Vienna 2006–present) and founding member of the Societas Celtologica Europaea (European Association of Celtic Studies scholars). His research interests are language variation and change in Old Irish and comparative Celtic linguistics. Research projects include a dictionary

of the Old Irish glosses in the Milan manuscript *Ambr. C301 infr.*, Lexicon Leponticum, and the ERC-funded project *Chronologicon Hibernicum*. His introductory handbook *Sengoídelc: Old Irish for beginners* (Stifter 2006) has been adopted for teaching Old Irish in universities worldwide.

Jürgen Uhlich is a lecturer in Early Irish language and literature at Trinity College Dublin. A monograph based on his Ph.D., entitled *Die Morphologie der komponierten Personnenamen des Altirischen*, appeared in 1993 (Uhlich 1993). Jürgen Uhlich has published on Early Irish and Celtic phonology and nominal morphology, the linguistic position of the earliest attested Celtic language Lepontic, Early Irish textual criticism as well as stylistics, most recently on the use of linguistic registers for stylistic purposes in the early Middle Irish text *Fingal Rónáin*. These areas also represent his ongoing research interests. He has furthermore prepared an edition of the Armenian translation of part of the acts of the *Ecumenic Concilium Ferrariense-Florentinum-Romanum* (1438–1445). He is currently working on a *Handbook of early Old Irish*, as well as on various linguistic and textual aspects of that linguistic period of Irish individually.

Christopher Yocum obtained his Ph.D. in Celtic Studies from the University of Edinburgh in 2009 for a thesis entitled *The literary figure of Fíthal*, which focused on the literary aspects of the early Irish judge Fíthal. His current research interests are in the application of semi-structured database and linked data concepts to the early Irish genealogical corpus. He has published articles in *Studia Celtica* and *Éigse*.

Overview of linguistic annotation

Linguistic examples quoted in the chapters are given interlinear glosses and English translations. The glossing conventions followed here are laid out in the following sections.

1 Glossing of Old Irish examples

Nouns are glossed with their translational equivalent and followed by the case (NOM, ACC, GEN, DAT) in subscript small capitals. Singular number is viewed here as default and is not glossed. Plural nouns are glossed with the tag PL, added after the case abbreviation following a full stop (e.g. NOM.PL).

(1) *feraib*
men$_{\text{DAT.PL}}$

(2) *geinti*
gentiles$_{\text{NOM.PL}}$

Adjectives are glossed with their translational equivalent and followed by case, number (SG, PL), and gender (MASC, FEM, NEUT) in subscript small capitals, each tag separated by a full stop.

(3) móir
big$_{\text{ACC.SG.FEM}}$

The definite article and other prenominal modifiers (such as quantifiers) are, generally speaking, glossed in the same way as an adjective. However, when the definite article is found immediately before a stressed demonstrative, no gender features are tagged since the demonstrative itself lacks clearly discernible gender features.

(4) a. *in* *fer*
the$_{\text{NOM.SG.MASC}}$ man$_{\text{NOM}}$
b. *in* *só*
the$_{\text{NOM.SG}}$ this$_{\text{NOM}}$

https://doi.org/10.1515/9783110680744-204

The unstressed demonstrative particles, *-sin* distal ('that') and *-so* proximal ('this') are glossed respectively as DIST and PROX. These tags are attached to the preceding item with the equals sign. Stressed demonstratives are tagged as nouns, as in (4b) above.

(5) a. *in* *fer-sin*
the$_{\text{NOM.SG.MASC}}$ man$_{\text{NOM}}$=DIST
b. *in* *fer-so*
the$_{\text{NOM.SG.MASC}}$ man$_{\text{NOM}}$=PROX

The stressed anaphoric pronoun, *suide* (in all case forms) is glossed with the tag ANAPH followed by case and number tags in subscript capitals with full stops between each tag type. Note that, as with nouns, singular is default and is not tagged. The unstressed anaphoric particle, which has the forms *side, sidi, ade, de, adi, di,* is only glossed with the tag ANAPH.

(6) a. *trisodin*
through=ANAPH$_{\text{ACC}}$
b. *achotlud* *adi*
his=sleep$_{\text{NOM}}$ ANAPH

Prepositional pronouns are glossed with the translational equivalent of the basic preposition followed by tags for person, number, gender, and case (in that order) in subscript small capitals. Tags for gender and case are separated from the tags for person and number with a full stop. The case tag is only used to disambiguate between the two possible cases (accusative and dative) governed a subset of prepositions which can govern both of these cases. If the preposition only ever governs one case, the case is not indicated in the glossing.

(7) a. *dóib*
to$_{\text{3PL}}$
b. *foir*
on$_{\text{3SG.MASC.ACC}}$
c. *for*
on$_{\text{3SG.MASC.DAT}}$

Verbs are glossed with their translational equivalent and followed by abbreviations in subscript small capitals for agreement, tense, mood, passive and relative (in that order) with a full stop between each abbreviation. The abbreviations

used are listed in (8). Note that indicative mood is here conceived of as the default and is not glossed.

(8) a. Tense: $_{\text{PRES}}$ (present), $_{\text{IMPF}}$ (imperfect), $_{\text{PST}}$ (past, only in past subjunctive), $_{\text{PRET}}$ (preterite), $_{\text{FUT}}$ (future).
b. Mood: $_{\text{SUBJ}}$ (subjunctive), $_{\text{CND}}$ (conditional), $_{\text{IMPV}}$ (imperative).
c. Passive forms are tagged $_{\text{PASS}}$; relative forms are tagged $_{\text{REL}}$.
d. Agreement: $_{\text{1SG}}$, $_{\text{2SG}}$, $_{\text{3SG}}$, $_{\text{1PL}}$, $_{\text{2PL}}$, $_{\text{3PL}}$.
e. The augment is tagged AUG or $_{\text{AUG}}$ (see below).

The sequence of glosses in verbs and examples of the method of glossing is given in (9). AUG has two positions. If it is the first preverb in the verbal complex it is treated as a PV (see below), consider (9a). If it is not the first preverb in verbal complex, it is glossed as in (9c).

(9) a. *ro·berthae*
AUG·bring$_{\text{3SG.PST.SUBJ.PASS}}$
b. *berthar*
bring$_{\text{3SG.PRES.SUBJ.PASS.REL}}$
c. *inroigrainn*
PV·persecute$_{\text{AUG.3SG.PRET}}$

For compound verbs, the lexical preverb is glossed separately as PV in capitals. Preverbs are separated from verbal roots by a raised dot in the glossing, even when the dot does not appear in the quoted example. Where present, infixed pronouns (glossed as 1SG, 2SG, 3SG_{MASC}, 3SG_{FEM}, 3SG_{NEUT}, 1PL, 2PL, 3PL) are inserted after the PV (or AUG) after a hyphen. If relevant the class type is added in parentheses in superscript afterwards (e.g. $3\text{SG}_{\text{NEUT(A)}}$, $3\text{SG}_{\text{NEUT(B)}}$, $3\text{SG}_{\text{NEUT(C)}}$). The hyphen is also used for the infixed relative, which is glossed REL, in prepositional relatives at after the preverbs *imm* and *ar*.

Consonant mutations play an important role in all Insular Celtic languages. In Od Irish, there are two prominent ones: lenition and nasalization. Lenition causes an initial stop to become a fricative; nasalisation causes initial voiceless stops to become voiced and prefixes a homorganic nasal to initial voiced stops and vowels. The mutations are glossed as superscript $^{\text{LEN}}$ and $^{\text{NAS}}$ respectively before the mutated form. Examples that follow these rules are given in (10).

(10) a. *as·beir*
PV·say$_{\text{3SG.PRES}}$

b. *at·beir*
PV-3SG$_{NEUT}$·say$_{3SG.PRES}$
c. *as·mbeir*
PV·^{NAS}say$_{3SG.PRES}$
d. *rondasaibset*
AUG-^{NAS}3SG$_{FEM}$·pervert$_{3PL.PRET}$
e. *immetét*
PV-REL·surround$_{3SG.PRES}$

Old Irish possesses a series of pronominal clitics that serve, roughly speaking, to emphasise items to which they cliticise. In traditional Irish grammar, these are called *notae augentes*. They are glossed with 1SG, 2SG, 3SG$_{MASC}$, 3SG$_{FEM}$, 3SG$_{NEUT}$, 1PL, 2PL, 3PL. These abbreviations are not in super/subscript. They are separated from the glosses for the stressed word with an equal sign (=) as in (11); see also below.

(11) *as·beir=som*
PV·say$_{3SG.PRES}$=3SG$_{MASC/NEUT}$

The example itself is presented using the editorial conventions of the edition cited. For example, if the edition does not use a raised dot to separate preverb from root, or a hyphen or equals sign to separate a *nota augens* from the verb, these are not inserted into the main text of the example. Punctuation is only inserted into the gloss as in (12).

(12) *asbeirsom*
PV·say$_{3SG.PRES}$=3SG$_{MASC/NEUT}$

In the gloss, an equal sign is used to separate an unstressed element from a stressed element (13), when the two are not separated by a space in the edition cited. A hyphen is used to separate an unstressed element from another unstressed element (14). A period is inserted between the words of translational equivalents where these consist of two or more words (15). An underscore is used between two possibly stressed items that are written without separation in the example (16).

(13) *isuidiu*
in=ANAPH$_{DAT}$

(14) a. *arní*
for-NEG
b. *donaibferaib*
for-the$_{\text{DAT.PL.MASC}}$=men$_{\text{DAT.PL.}}$

(15) *mórabba*
great.cause$_{\text{ACC}}$

(16) *ísíu*
DEICT_this$_{\text{DAT}}$

Note that (16) shows that the deictic particle í is glossed as DEICT. The negative particles are glossed NEG (main clause *ní*), NEGSUB (non-main clause *na/nach/nad*) in subordinate non-relative clauses and NEGREL in relative clauses.

2 Glossing of Brittonic examples

The glossing of Brittonic examples is somewhat different from the glossing of Old Irish. These differences are exemplified below.

Nouns and adjectives are glossed with their translational equivalent only.[1]

(17) a. *gwin*
wine
b. *riuedi*
numbers

(18) *margh uskis*
horse swift

The definite article is glossed as DEF.

(19) *'r llys*
DEF court

1 Very occasionally, subscript small capital $_{\text{PL}}$ is used to disambiguate a plural form of an adjective from a non-plural form (e.g. Welsh *eraill* is glossed other$_{\text{PL}}$). Certain numerals have feminine and masculine forms. These are distinguished with subscript small capital $_{\text{FEM}}$ and $_{\text{MASC}}$, (e.g. *tri* three$_{\text{MASC}}$ vs *tair* three$_{\text{FEM}}$).

All pronouns in Brittonic are tagged with the appropriate agreement tag (1SG, 2SG, 3SG, 1PL, 2PL, 3PL) and, if necessary, the following tags in subscript capitals: MASC, FEM, POSS (possessive), INFX (infixed) INTS (intensifier), REFL (reflexive).

(20) a. *y* *penn*
3SG$_{\text{MASC.POSS}}$ head
b. *a* *'e* *lladwn* *ef.*
PTCL 3SG$_{\text{MASC.INFX}}$ kill$_{\text{1SG.SUBJ.IMPF}}$ 3SG$_{\text{MASC}}$
c. *dy hun*
2SG$_{\text{INTS}}$
d. *dy hun*
2SG$_{\text{REFL}}$

All demonstratives in Brittonic are tagged as either DIST (distal) or PROX (proximal).

(21) a. *henna*
DIST
b. *an* *den* *ma*
DEF man PROX
c. *hynny*
PROX

Verbs are glossed with their translational equivalent and followed by abbreviations in subscript small capitals for agreement, tense, mood, and impersonal (in that order) with a full stop between each abbreviation. The abbreviations used are listed in (22). Note that indicative mood is here conceived of as the default and is not glossed.

(22) a. Agreement: 1SG, 2SG, 3SG, 1PL, 2PL, 3PL.
b. Tense: PRES (present), PRET (preterite), FUT (future), IMPF (imperfect), PLPF (pluperfect), HAB (habitual).
c. Mood: SUBJ (subjunctive), COND (conditional), IMPV (imperative).
d. IMPS (impersonal)
e. The perfective particle *re, ry, 'r* (*etc.*) is tagged PERF.

The sequence of glosses in verbs and examples of the method of glossing is given in (23).

(23) a. *ledy*
kill$_{2SG.PRES}$
b. *deuthant*
come$_{3PL.PRET}$
c. *lladwn*
kill$_{1SG.IMPF.SUBJ}$
d. *wnathoed*
do$_{3SG.PLPF}$
e. *bythynt*
be$_{3PL.HAB}$

The particle *ym-* (also spelled *em-*) is glossed PV. This is separated from verbal roots by a raised dot in the glossing. Infixed pronouns (glossed as 1SG$_{INF}$, etc.) are separated from the verb and supporting particles by whitespace. Examples that follow these rules are given in (24).

(24) a. *ym·dodant*
PV·melt$_{3PL.PRES}$
b. *re gowsys*
PERF·speak$_{3SG.PRET}$
c. *ny* *'s* *gwna* *e hun*
NEG 3SG$_{MASC.INF}$ make$_{3SG.PRES}$ 3SG$_{MASC.INTS}$

Other verb-related glosses are: VN (verbal noun), PST-PTCPL (past participle), PTCPL (participle), all subscript small capitals.

Negative particles are glossed NEG, with subscript SUB used for the subordinate negative, where necessary. The predicative particle (*yn* in Welsh) is glossed PRED. The progressive particle (*ow* in Cornish) is glossed PROG. Other particles are glossed PTCL.

3 List of abbreviations

1	1st person
2	2nd person
3	3rd person
A	Class A pronouns
ACC	Accusative
ANAPH	Anaphor
AUG	Augment
B	Class B pronouns

C	Class C pronouns
CND	Conditional
DAT	Dative
DEF	Definite
DEICT	Deictic particle í
DIST	Distal Demonstrative
FEM	Feminine
FUT	Future
GEN	Genitive
HAB	habitual
IMPF	Imperfect
IMPS	Impersonal
IMPV	Imperative
INF	Infinitive
INFX	Infix
INTS	Intensifier
LEN	Lenitition
MASC	Masculine
NAS	Nasalization
NEG	Negation
NEUT	Neuter
NOM	Nominative
PASS	Passive
PERF	Perfect
PL	Plural
PLPF	Pluperfect
POSS	Possessive
PRED	Predicative Particle
PRES	Present
PRET	Preterite
PROG	Progressive
PROX	Proximal Demonstrative
PST	Past (Subjunctive)
PST-PTCPL	Past passive participle
PTCPL	Participle
PV	Preverb
REFL	Reflexive
REL	Relative
SG	Singular
SUB	Subordinate (Negative)
SUBJ	Subjunctive
VN	Verbal Noun

Elliott Lash, Fangzhe Qiu, and David Stifter

Introduction: Celtic Studies and Corpus Linguistics

1 Background to the volume

This volume is a collection of eleven chapters that showcase the state of the art in corpus-based linguistic analysis of the old, middle and early modern stages of Celtic languages (specifically, Old and Middle Irish, Middle Welsh, and Cornish). The contributors offer both new analyses of linguistic variation and change as well as descriptions of computational tools necessary to process historical language data in order to create and use electronic corpora. On the whole, the volume represents a platform for the exploration of corpus approaches to morphosyntactic variation and change in the Celtic languages and, for the first time, situates Celtic linguistics in the broader field of computational and corpus linguistics.

These chapters were originally prepared for lectures hosted by the Chronologicon Hibernicum project (ChronHib), an ERC-funded project at Maynooth University, Ireland (ERC Consolidator Grant 2015, H2020 #647351). The lectures occurred at three separate workshops (December 15, 2016, April 4, 2017, October 13–14, 2017), which brought together an international group of researchers with various backgrounds to help the ChronHib team gain insight into preparing linguistically marked-up text for statistical research on language variation in Old Irish. At the first event, all aspects of corpus building and use, such as morphological tagging, syntactic parsing and maintenance and sustainability of online databases, were discussed. In subsequent events, two main themes emerged: first, the necessity of developing computational tools such as morphological taggers/analysers and lemmatisers, and second, that careful use of corpora with a focus on new search queries yields progress on previously intractable problems of Celtic morphosyntax.

2 ChronHib and CorPH

The overall goal for ChronHib is to develop a statistical methodology of linguistic dating in order to more precisely date the diachronic development of the Early Irish language (Old Irish: seventh to ninth century, Middle Irish: tenth to twelfth century) and thereby to predict the age of the large number

https://doi.org/10.1515/9783110680744-001

of anonymous, dateless Irish texts. In many ways, too, the early stages of Brittonic languages present the same problems of anonymous, as yet undated text (Rodway 2013). In traditional studies of both Goidelic and Brittonic material, linguistic dating has typically been a matter of philological and linguistic analysis of manually curated data. ChronHib aims at advancing the methods used for linguistic dating of Early Irish by contributing to a chronologically more precise description of linguistic variations and by employing corpus linguistic and advanced statistical methods. It also endeavours 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 (see Qiu et al. 2018 for a more in-depth discussion of ChronHib).

Essentially, ChronHib will produce a new linguistically tagged corpus of Old Irish texts. This corpus, called the *Corpus Palaeohibernicum* (CorPH, Stifter et al. 2015–) is in the development stage and will soon be freely accessible online. It will, firstly, unify some of the existing resources for the study of Old Irish texts under one annotation scheme, and secondly, expand the amount of electronic materials by digitising and annotating data that have only been available previously in printed media or manuscripts. Scholars working on Old Irish, for example, have, until now, mainly relied on the data found in the two-volume printed edition of *Thesaurus Palaeohibernicus* (Thes. = Stokes and Strachan 1901–1910). The existing digital resources for medieval Irish texts come in a variety of forms: annotated lexicons, digital glossaries, text with XML markup, treebanks, and fully digital dictionaries. For extensive discussion of some of these materials, see Griffith, Stifter, and Toner (2018). These heritage data together constitute the corpus on which the contributions in this volume are based, and a brief description of them is pertinent here.

The main online dictionary of Early Irish is eDIL (Toner et al. 2019). It enables research into semantic, morphological, and syntactic usage of Irish lexemes in sources written between the seventh century and 1700. There are, in addition, two major digital collections of early Irish texts: the *Corpus of Electronic Texts* (*CELT*) hosted by University College Cork (Färber 2012) and the *Thesaurus Linguae Hibernicae* (*TLH*) hosted by University College Dublin (Kelly and Fogarty 2006–2011). These corpora consist of analytically and structurally XML-marked up texts following the TEI guidelines. The usefulness of these textual resources for the corpus-linguist is only indirect, since no linguistic information is tagged. A prominent treebank is the *Parsed Old and Middle Irish Corpus* (Lash 2014a), a UPenn-style syntactically tagged treebank of fourteen Old Irish texts. The two online annotated lexicons are the *Milan Glosses* database (Griffith and Stifter 2013) and the *Priscian Glosses* database (Bauer 2015; see also Bauer, Hofman, and Moran 2018). These are fully annotated

for morphological and lexical information. Griffith and Stifter's (2013) database consists of around 50,000 morphologically and POS-tagged tokens from the Old Irish glosses in the Milan manuscript *Ambr. C301 infr.* (Ml.). Bauer's (2015) database consists of around 20,000 morphologically and POS-tagged tokens from the Old Irish glosses in several manuscripts of Priscian's *Institutiones Grammaticae*, with the St Gall *Stiftsbibliothek manuscript 904* (Sg.) containing the most extensive collection of these glosses. These two databases, along with the *Lexicon of the Old Irish glosses in the Würzburg manuscript of the Epistles of St. Paul* (Wb.; Kavanagh 2001, available in print and .pdf formats), have been the catalyst for much research into linguistic variation in Old Irish over the past eighteen years.

The above databases (Ml., Sg.) and lexicon (Wb.) were used by most of the contributors in the present volume who studied variation in Old Irish in contemporary (eighth to ninth century) manuscripts. Moreover, many of the texts discussed in Liam Breatnach's and Christopher Yocum's contributions can be found in the CELT and TLH corpora. The Ml. and Sg. databases have now been incorporated into CorPH and stand beside other resources specifically made for CorPH such as the *Minor Glosses* database (Lash 2018), the *Annals of Ulster* database (Qiu 2019), and the *Poems of Blathmac* database (Barrett 2018a) In total, CorPH has over 120,000 fully annotated tokens of Old Irish text in various genres (glosses, annals, poetry, chief among them) and will allow researchers easy access to a large amount of data for research on linguistic variation. Some chapters in this volume (for example, Elisa Roma's and Theodorus Fransen's) have already made use of data from CorPH.

For the other well-attested medieval Celtic language, Middle Welsh (c. 1150–1500), authoritative editions have long served as the standard corpus for scholars. Meanwhile, two online, searchable corpora have been published, covering the majority of prose texts surviving from before 1425: *Rhyddiaith Gymraeg o Lawysgrifau'r 13eg Ganrif* (Isaac et al. 2013) and *Rhyddiaith Gymraeg 1300–1425* (Luft, Thomas, and Smith 2013). These form the basis of Britta Irslinger's investigation in this volume, and a more detailed description can be found in that contribution. The late medieval and early modern period of the Welsh language is represented by the *Corpws Hanesyddol yr Iaith Gymraeg 1500–1850* (Willis and Mittendorf 2004), which contains about 420,000 words from 30 texts in a variety of genres. However, these corpora have not been linguistically tagged and therefore their usefulness is somewhat limited. The contribution by Marieke Meelen aims to tackle this lacuna by developing tagging methods for part of the prose corpora mentioned above. The last medieval Celtic language dealt with in this volume, Cornish in its middle (c. 1200–1600) and late (c. 1600–1750) phases, survived mainly in versified religious plays and translated works, scholarly

editions of which constitute the corpus for the analysis in Joseph Eska and Benjamin Bruch's contribution.

3 Overview of themes

Digital corpora for medieval Celtic languages have certainly become a central part of the field of Celtic Studies in recent years but fully annotated corpora are still few in number and the application of computational linguistic methods in the analysis of Celtic languages is in its infancy. These languages represent a new frontier in the development of natural language processing tools, in part because they pose special challenges, such as complicated inflectional morphology with non-straightforward mappings between lemmata and attested forms, highly variable orthography, and initial consonant mutations. With so much data available in non-electronic form as the result of previous work and ongoing efforts to convert these data to computer-readable format, it is not surprising to find that the contributors employ both available digital corpora and printed editions or manuscripts in their research, and that quantitative studies are more often conducted in a data-based or data-inspired rather than data-driven manner. This approach shows great potential in revealing hitherto subtle generalisations over various aspects of medieval Celtic languages.

A significant aspect of the volume is that the quantitative studies all deal with aspects of syntactic structure, a subsection of the grammar of medieval Celtic languages (Irish in particular) that has suffered relative neglect, in favour of investigations focusing on phonology and morphology. Happily, more work on syntax has appeared since Isaac (2003) gave a short survey of the few works in the field and pronounced a handbook of Old Irish syntax to be a desideratum. Much of the work of the last decade and a half (e.g. García-Castillero 2013; Griffith 2008; Lash and Griffith 2018; Roma 2014) draws directly on the increasing availability of searchable corpora that enable easy access to the fundamental dataset. This explosion in research is set to continue with the development of CorPH. Bringing the results produced by central scholars participating in this endeavour together in one place emphasises the potential that corpus approaches have in aiding research and underlines many points in need of further investigation.

With its concentration on computational corpus linguistics and morphosyntactic data from historical language stages, this volume is a first in the discipline of Celtic Studies, which has been mainly focussed on traditional philological work such as the editing of texts and literary/historical explication of these

texts. Additionally, it contrasts with and complements other recent volumes of interest to scholars working in Celtic Studies, such as *Formal Approaches to Celtic Linguistics* (Carnie 2011), *Linguistic and Philological Studies in Early Irish* (Roma and Stifter 2014), the proceedings of the fourth *International Congress of Celtic Studies, held in Maynooth University, 1–5 August 2011* (Breatnach et al. 2015), and *Centres and Peripheries in Celtic Linguistics* (Bloch-Trojnar and Ó Fionnáin 2019). While each of these volumes consist of chapters analysing various stages of the Goidelic and Brittonic languages, very few use corpus data or deal with problems of corpus building. Moreover, many of these contain chapters that are more philological, historical, or literature-oriented than strictly linguistic in nature. The present volume, in contrast, reflects the increasing awareness of the usefulness of corpus data in Celtic linguistics, and its contributions show how corpora of Celtic languages can be most effectively constructed and exploited. In the meantime, scholars who focus mainly on philology should still find many of the chapters interesting, as they contribute to our knowledge of the grammars of medieval Celtic languages from fresh perspectives. It is also hoped that chapters such as Marieke Meelen's and Theodorus Fransen's, which showcase the development and testing of new computational tools for Celtic language data will also appeal to linguists in general, especially those who are interested in diachronic linguistic changes, computational linguistics, and corpora of historical languages.

4 Description of chapters

The volume is divided into two thematically distinct but related parts. Part one consists of four chapters dealing with the design and creation of corpora for historical languages generally and Celtic languages in particular. Part two consists of seven chapters that are broadly united by the theme of description and qualitative/quantitative analysis of linguistic data derived from the available corpora of medieval Celtic languages. The division into two main parts is motivated by thematic concerns, since the contributions fall into two general groups. There are, firstly, detailed technical discussions of corpus construction, automatic annotation tools, and clustering methods (Marius Jøhndal, Theodorus Fransen, Marieke Meelen, and Christopher Yocum's chapter), and secondly, primarily corpus-based analyses of particular phenomena (Liam Breatnach, Carlos García-Castillero, Jürgen Uhlich, Elisa Roma, Aaron Griffith, Joseph Eska and Benjamin Bruch, and Britta Irslinger's chapter). The first part of the book is therefore, roughly speaking, practical with its concentration on computational research tools and methods, while the second is analytical in focus.

Within each part of the book, chapters are themselves grouped thematically. Part one begins with two chapters (by Marius Jøhndal and Marieke Meelen, respectively) that originate from discussions at the first and second ChronHib workshops about the building and sustainability/maintenance of linguistically annotated corpora. Additionally, as a description of a new Welsh treebank, Meelen's chapter responds to some of the concerns about the need for better ways of doing research on problems of Celtic syntax, as was expressed by participants at the second and third ChronHib workshops. The next two chapters in part one concentrate on the creation and use of computational tools in order to analyse particular aspects of the Old Irish corpora (verbal morphology in Theodorus Fransen's chapter and stylistic clustering in Christopher Yocum's chapter).

Part two begins with two chapters (by Liam Breatnach and Carlos García-Castillero, respectively) that investigate the diachronic syntax and morphology of pronouns and demonstratives in Old Irish The following three chapters (by Elisa Roma, Jürgen Uhlich, and Aaron Griffith, respectively) are all united through their investigation of grammaticalised consonant mutations in Old Irish, whether in the context of relative clauses (Griffith and Uhlich) or after nominals (Roma). The final two chapters in part two (by Joseph Eska and Benjamin Bruch on the one hand and Britta Irslinger on the other) deal with some syntactic phenomena in the Brittonic languages.

4.1 Description of Part 1

Marius Jøhndal's "Treebanks for historical languages and scalability" presents both a general overview of the motivations for and practice of corpus building as well as a detailed overview of the PROIEL family of treebanks. This group of treebanks includes annotated texts from older Indo-European languages and is one of the most ambitious recent corpus-related projects for these languages. It includes the original core, the PROIEL (Pragmatic Resources in Old Indo-European Languages) itself, which is a corpus of New Testament texts in Ancient Greek, Latin, Classical Armenian, and Gothic, as well as some other texts in some of these languages. Additionally, the PROIEL family also includes the ISWOC Treebank, consisting of texts in Old English and Old Romance (Spanish, Portuguese), and the TOROT database with texts in Old Slavic (Old Church Slavonic, Old Russian). One of the goals of the chapter is the introduction of a new interface for browsing and searching the PROIEL Treebank and related treebanks called *Syntacticus* (http://syntacticus.org). This expansion of the PROIEL family of treebanks increases its visibility and is a crucial way of

achieving long-term maintenance. It is also an exemplary open-source infrastructure that can be used for future projects. The chapter is therefore programmatic and practical, since the kinds of technical, linguistic, and manpower related challenges it describes serve as both a guideline to best practice and an inspiration for future research on Celtic languages. Although the chapter does not discuss Celtic languages in particular, in many respects it sets the tone for the volume since many of the issues mentioned in it, being characteristic of less-resourced historical languages, will be familiar to scholars of medieval Celtic languages and it is hoped that the chapter may serve as a call to collaboration.

"Annotating Middle Welsh: POS tagging and chunk-parsing a partial corpus of native prose" by Marieke Meelen demonstrates the process of creating an annotated corpus of some Middle Welsh native prose (as against translated works), and the challenges and potentials of building such a corpus. The corpus contains only literary narratives and some law texts at present but will be extended to other genres and registers. Digitalised texts were pre-processed with punctuation and tokenisation, which was done automatically by a POS tagger and a Memory-Based Tagger. The text was then marked up with a simplified version of the TEI P5 header. The author adopts the UPenn annotation scheme modified with Welsh-specific tags that enable further queries concerning agreement patterns and change in Information Structure. A Memory-Based Tagger assigns morpho-syntactic tags to tokens automatically and a modified rule-based chunk-parser is deployed to annotate syntax and information structure. This chapter presents the first systematic approach to annotating historical Welsh, and the corpus it describes ultimately aims to provide a starting point to build a fully annotated Welsh historical treebank.

In "Automatic morphological analysis and interlinking of historical Irish cognate verb forms", Theodorus Fransen describes a computational approach to understanding how the Irish verbal system develops diachronically. The author's major contribution is to propose a morphological analyser for Old Irish verbs and to discuss ways this analyser can be incorporated into a framework of computational resources for various stages of Irish. This proposal dovetails with Jøhndal's and Meelen's chapters in dealing with ways of expanding the current computational toolset for a historical language (specifically historical stages of Irish) and in its concerns with scalability. These concerns are reflected in his detailed investigation of the challenges encountered by a methodology that incorporates finite-state morphology as it applies to Old Irish. The challenges he details are two-fold. The first challenge has to do with word and morpheme division as encountered in "real" text, i.e. editions or manuscript transcriptions. In many cases, multiple morphemes may be written as a concatenated string, resulting in the need to find a way to encode licit combinatorial possibilities of multiple

morphemes. This is a so-called *generation* problem, where generation means the ability of the analyser to generate all and only the licit inflected forms of any given stem. In other cases, whitespace is found between morphemes leading to potential parsing ambiguities since the analyser is word-based (where a word is understood to be an element between whitespace). This is a so-called *analysis* problem, which may result in the wrong morphological tag being assigned to any given string. The second challenge has to do with the complex interaction between phonology (especially stress) and morphology in Old Irish since stress alternations can result in syncope and the presence or absence of palatalisation of stem-final and ending-initial consonants. These challenges impinge on the choices made for implementing the finite-state transducer. For instance, does one rely on a strictly rule-based approach to specify certain licit combinations and handle stem variants induced by stress alternations, using "flag" morphemes or upper-level filters for instance to deal with the generation problem? Or does one hard-code (i.e. list) such stem variation or parts of paradigms? Fransen carefully weighs the advantages of different approaches in order to ensure the applicability of his analyser. He also envisions a fully functioning POS-tagger suitable for both Old and Middle Irish by making some suggestions for allowing interoperability of resources, especially between his morphological analyser and Dereza's (2018) Old Irish lemmatiser.

Christopher Yocum's chapter "Text clustering and methods in the *Book of Leinster*" uses machine-learning techniques to cluster the texts in the *Book of Leinster* (LL), and tries to identify the reason for the clustering. The author extracts individual texts from the electronic edition of LL, tags the function words and calculates the frequency of function words in each text. The frequencies are then turned into a matrix of vectors, which goes through the *k*-medoids algorithm, subject to normalisation and "Principal Component Analysis". The result is a clustering scatter plot. The clustering can be caused by the variables of author, scribe or genre, and these three factors are tested in turn. The result suggests that authorship is the main factor in clustering, and that the traditional ascriptions to certain authors do not fit the clustering and may need to be revised. The methods used are innovative within Celtic Studies and contrast with the traditional philological approach to text clustering. The chapter is a useful addition to the large body of work on the history of the manuscript and the clusters of text reported on deserve further investigation. If specific linguistic usages can be associated with particular clusters, this may be useful for the study of idiolect/style at particular periods.

4.2 Description of Part 2

In "The demonstrative pronouns in Old and Middle Irish", Liam Breatnach uses a corpus of Old Irish verse texts that are largely available online in TLH and CELT. The author first observes that there is a split between the unstressed enclitic demonstrative particles *-sin* 'that', *-so/se* 'this' and their stressed pronominal variants, *sin* 'that', *só/sé* 'this' (dative *sund/síu*). The rest of the chapter deals with a diachronic investigation of the morphophonology, syntax, and semantics of the stressed demonstrative pronouns. The results of this investigation map the distribution of demonstratives according to four main features: syntactic function, singular/plural number, inanimate/animate reference and period (i.e. Old versus Middle Irish). The main contribution of the chapter is that it highlights subtle differences between Old and Middle Irish usages. First, while the stressed demonstratives on their own (without the addition of the particle *í*) could be construed as plural in both Old and Middle Irish, plural reference was very restricted in Old Irish, but much expanded in Middle Irish. Specifically, plural reference is found in Old Irish when the demonstrative acts as a subject of a copular sentence and in later Old Irish as the complement of an agreeing preposition. Middle Irish allows plural reference in some other contexts. Second, demonstratives with inanimate and animate reference are likewise found in both Old and Middle Irish, but animate reference in Old Irish once again is restricted to subjects of copular sentences whereas it is found in other contexts in Middle Irish. The chapter closes with some discussion of the possibility that the independent, personal pronoun *sé* 'he' developed during Middle Irish from the demonstrative *sé* in contexts where it had animate reference.

Carlos García-Castillero's chapter is titled "Paradigmatic split and merger: The descriptive and diachronic problem of Old Irish class B infixed pronouns". This contribution replaces García-Castillero's lecture "Synonymy (a^N / *ané* 'that (what)', a^N / *inta(i)n* 'when') and homonymy (a^N 'that (what)' and a^N 'when') in the Old Irish glosses" presented at the third workshop, because the author had already submitted the lecture for publication elsewhere. The contribution in this volume explains the diachronic origin of the Old Irish class B infixed pronouns, which are used in a declarative clause after pretonic lexical preverbs of the structure (-)VC-. The author firstly clarifies the relevant notions in Old Irish (clause types, verbal complex, phonotactic structure of preverbs, etc.), and then illustrates the use of non-third person infixed pronouns with instances collected from the corpus of the contemporaneous Old Irish glosses. This corpus-based approach yields the interesting observation that, in the language of the contemporaneous Old Irish texts, non-third person infixed pronouns are much less regular than the third person infixed pronouns in making a distinction between declarative and relative forms,

especially when the lexical preverb after which the infixed pronoun appears is of type (-)VC-. Such asymmetry in distribution between the persons raises a question, which, in the author's opinion, is directly related to the diachronic origin of the class B infixed pronouns. The author argues that class B infixed pronouns arose to distinguish a verbal complex with a third person singular masculine or neuter infixed pronoun in a declarative clause from a complex without an infixed pronoun in a relative clause. More specifically, a process of morphological split in the original class C paradigm has given rise to two forms in the third persons, and tentatively in the other persons.

Elisa Roma presents her findings on the distribution of nasalisation after nominals in Old Irish glosses in "Nasalisation after inflected nominals in the Old Irish glosses: Evidence for variation and change", where her main interest lies in the possibility of mapping variation in nasalisation to chronological or diatopic criteria. All instances of nasalisation after nominals from four Old Irish corpora of glosses have been collected (Wb., Ml., and Sg. and the *Minor Glosses Database*). The phonetic contexts for nasalisation are categorised, as well as the word class of the nasalising/nasalised word. The frequency of nasalisation in each combination of phonetic context and word class has already been reported in Roma (2018a). Firstly, the data show that the absence of nasalisation after inflected nominals in Old Irish cannot be due merely to the loss of a nasal consonant in consonant clusters. Secondly, individual texts show different frequencies of nasalisation in the same context. The variation between Old Irish texts in nasalisation after inflected nominals suggests not only diachronic strata but also probable regional differences that led to later developments in Modern Irish and Scottish Gaelic. The chapter is comparable to other corpus-based investigations of morphophonology, such as Griffith (2016a) and Lash (2017a). Together with these papers, Roma's chapter is illustrative of the impact lexicons and corpora have had on Celtic linguistics.

In "On the obligatory use of a nasalising relative clause after an adjectival antecedent in the Old Irish glosses", Jürgen Uhlich uses a corpus consisting of the main Old Irish glosses (Wb., Ml., Sg.) to explore the extent to which adjectives having a modal adverbial reading must be followed by a nasalising relative clause in cleft sentences (e.g. *arndip maith nairlethar a muntir* 'so that he may well order his household', lit. 'that it may be good how he orders'). The author argues that, save for some well-defined exceptions, the nasalising relative clause is an absolute prerequisite of this construction. His approach is at once quantitative, since he has systematically and exhaustively collected all instances of modal adjective cleft sentences from the glosses and studied their distribution, and qualitative, since he also carefully establishes and describes the varying types of "exceptions" to the generalisation. The exceptions to the

generalisation include (a) cases in which the verb in the clause following the adjective has an object marked with a class A or B infixed pronoun, (b) instances of mixed antecedents in coordination where the antecedent farthest from the embedded clause is the modal adjective, (c) clauses involving what Uhlich terms "syntactic raising", essentially multiple dependencies, where the modal adjective and another constituent simultaneously act as the antecedent to the embedded clause, and (d) some possibly innovative instances of leniting rather than nasalising relative clauses. The paper is an important contribution to a long-standing debate in Old Irish studies dealing with the rather complex syntax of relative clauses and its conclusion that a nasalising relative clause is an essential component in a modal adjective cleft revises the previous consensus that nasalising relative clauses were optional across much of the domain in which they could be used.

In Aaron Griffith's chapter, "The 'Cowgill particle', preverbal *ceta* 'first', and prepositional cleft sentences in the Old Irish glosses", he connects what he calls "three seemingly unrelated" phenomena: the phonological shape of the adverbial preverb *ceta* 'first', evidence for the so-called Cowgill Particle (**eti*), and the usage of relative verbs in PP-clefts. The author investigates both the first and second vowel in *ceta* using a combination of a quantitative corpus-based approach and a qualitative comparative approach. In his discussion of the variation in the initial syllable of *ceta* (attested as both *ceta* and *cita*), he shows that the usage of the *i*–variant increases over time. He then argues that the final vowel of *ceta*, together with the final vowel of the preverb *ocu* (in *ocu-ben*) could provide further, previously unexamined, evidence for the Cowgill Particle, if the initial vowel of **eti* was not elided after preverbs ending in *u* (i.e. **kintu-eti*, not **kintu-ti* > *ceta*, **onku-eti*, not **onku-ti* > **ocu*). Because the preverb *ceta* is predominately found in relative clauses, where the Cowgill Particle would in fact not be expected, the paper then shifts to a discussion of two examples in which a verb containing *ceta* is arguably non-relative. These two examples are both prepositional cleft sentences (e.g. *ar is do thabirt díglae berid in claideb sin* 'for it is to wreaking revenge that he carries that sword'), where a non-relative verb typically follows the prepositional phrase (PP). The author surveys the evidence for PP-clefts in the corpus of glosses and shows that, despite the general rule, the Milan Glosses have innovative relative verbs after the PP. While this leaves the status of the two examples containing *ceta* uncertain (they could either be non-relative, and therefore evidence for the Cowgill Particle, or relative), the chapter is, like Uhlich's, a useful contribution to the perennial debate on the syntax of cleft sentences and relative clauses in Old Irish.

Britta Irslinger, in "The functions and semantics of Middle Welsh X *hun*(*an*): a quantitative study", uses two untagged corpora of Middle Welsh – *Rhyddiaith*

Gymraeg 1300–1425 / Welsh Prose 1300–1425 and *Rhyddiaith y 13eg Ganrif: Fersiwn 2.0* – to investigate an innovative usage of the collocation *X hun*(*an*) (where X is a possessive pronoun) as a reflexive pronoun in Middle Welsh. The author shows that the collocation *X hun*(*an*) was generally used as an intensifier in the corpora, in a manner similar to English *myself* in *I saw him myself,* but there is some evidence of its grammaticalisation as a reflexive pronoun. This new function of *X hun*(*an*) appears in fourteen instances out of a total of 1908 unique tokens of *X hun*(*an*), where it is used instead of the usual reflexive markers, the verbal prefix *ym-* or plain pronouns. The fourteen examples of reflexive usage come from translation literature, but it does not appear that the collocation *X hun*(*an*) corresponds to any particular intensifier marker in the base language. This suggests that the examples display a real innovation in Welsh grammar. The study is part of an ongoing effort (see references cited in the chapter) to understand the expression of reflexivity, reciprocal action, and middle voice in Welsh and also contributes to the debate over the extent to which English *-self* as an expression of reflexivity arose as the result of contact with Welsh. According to the author, the use of *-self* as a reflexive in English expanded from the mid-twelfth to the seventeenth century. Although this is not explicitly stated by the author, the fact that there are so few examples of *X hun*(*an*) used as a reflexive before 1425, i.e. after the first signs of the innovation in English, could suggest that the contact with Welsh was not the only factor in the development of *-self*.

In "Prolegomena to the diachrony of Cornish syntax", Joseph Eska and Benjamin Bruch discuss the diachronic development of the configuration of the Cornish affirmative root clause with comparison to other Brittonic languages. Since verbal sequences do not occur in Old Cornish, examples from Old Welsh and Old Southwest Brittonic, showing VSO and V2 orders, are quoted, with the assumption that these languages behaved similarly to Cornish. The affirmative root clauses in Middle Welsh and Middle Breton are generally V2, and surface V2 (along with V3) is also found in Middle Cornish. The authors then analyse the architecture of the left periphery and the preverbal Object DP, pointing out that the exceptions to V2 in Middle Cornish are caused by metrical considerations overriding the grammar, and despite the corpus of Middle Cornish being composed largely of verse, the Middle Cornish affirmative root clause was V2 of the "relaxed" type. The authors then examine the corpus of Late Cornish texts and find that these are of dubious evidential value because the corpus is very small and consists of translations by a native speaker and texts by non-native speakers.

Part 1: **Corpus tools for historical Celtic linguistics**

Marius L. Jøhndal

1 Treebanks for historical languages and scalability

1 Introduction

Historical linguistics, whether synchronic or diachronic, is by definition based on corpora. Since we do not have access to the intuitions of native speakers we can only test linguistic hypotheses about historical languages by systematically collating information from our corpus of texts.

For questions that typically concern linguists, this often means identifying every occurrence of a particular phenomenon in the corpus, analysing, classifying and counting the occurrences and then using this for testing hypotheses about the structure of the language. This can be done manually, but this is time-consuming and error-prone. As Haug (2015) points out, while reading the text and manually collating information from it is essential for *hypothesis formation* it is much less useful for *hypothesis testing*. Even if the text is in electronic form, it is easy to overlook an example, record it incorrectly or fail to apply test criteria consistently over time.

This paper focuses on treebanks, which are corpora that have been annotated with morphosyntactic information so that we can extract linguistic structures like 'verb with an accusative noun'. High-quality treebanks for a range of historical languages now exist and are widely used in historical linguistic research. This includes treebanks that follow the Penn-style of annotation, e.g. the Penn-Helsinki Parsed Corpus of Middle English (Kroch and Taylor 2000), the Penn-Helsinki Parsed Corpus of Early Modern English (Kroch, Santorini, and Delfs 2004), the Penn-Helsinki Parsed Corpus of Modern British English (Kroch, Santorini, and Diertani 2016), the Tycho Brahe Parsed Corpus of Historical Portuguese (Galves and Britto 2002) and the Icelandic Parsed Historical Corpus (Wallenberg et al. 2011), as well as dependency-based treebanks, e.g. the Index Thomisticus (Passarotti 2007), the Ancient Greek and Latin Dependency Treebanks (Bamman and Crane 2011; Celano, Crane and Almas 2014), the PROIEL Treebank (Haug and Jøhndal 2008, Haug, Eckhoff et al. 2009), the ISWOC Treebank (Bech and Eide 2014) and the TOROT Treebank (Eckhoff and Berdičevskis 2015).

A key challenge in building treebanks for historical languages is lack of resources. Funding is limited and there are few existing computational language resources like taggers and parsers available. At the same time, the task is complex and experts on the language have to devote a significant amount of time to

https://doi.org/10.1515/9783110680744-002

the annotation task. This comes on top of the complexity of designing a suitable annotation scheme that balances the desire to capture philological and linguistic detail with an approach that is reliable, scalable and technically feasible.

A key motivator behind treebank efforts is to facilitate reuse of resources and to provide access to large data sets that make hypothesis testing robust and encourage replication of published research, but as funding for construction of a treebank tends to be tied to a time-limited research project, it is challenging to fulfil such long-term aspirations and achieve scale and long-term consistency.

This paper describes these challenges in the context of the PROIEL, ISWOC and TOROT treebanks, and how this has motivated efforts to use automated tools like taggers and parsers to scale the annotation process. The paper also describes *Syntacticus* (http://syntacticus.org), which now serves as a shared front-end for PROIEL, ISWOC and TOROT, but whose long-term aim is to integrate automated taggers and parsers with our existing annotation tools and offer this as an open infrastructure platform that can be used by researchers working on other less-resourced, historical languages within the Indo-European family, such as the Celtic languages.

Section 2 briefly introduces the PROIEL, ISWOC and TOROT treebanks and some key properties of the annotation scheme. Section 3 describes the challenges involved in maintaining these treebanks, expanding them and making them accessible for researchers, and how this has motivated us to set up Syntacticus. Section 4 describes in more detail current efforts aimed at evaluating how the annotation process can be scaled using automated taggers and parsers.

2 The PROIEL, ISWOC and TOROT treebanks

The PROIEL-family of treebanks currently includes the PROIEL, ISWOC and TOROT treebanks. Together they contain text samples from a number of old Indo-European languages (see Table 1) which, when consolidated into one treebank, contains around one million words that have been lemmatized, morphologically analysed and annotated with syntactic dependencies.

The original PROIEL Treebank stems from a research project called *Pragmatic Resources in Old Indo-European Languages* at the University of Oslo (2008–2012), which was set up to study information packaging in ancient Indo-European languages. A major part of this was to compile a treebank containing the New Testament in its original and translations, as the New Testament is a natural

Table 1: Languages and token counts in the PROIEL Treebank release 20180408, the TOROT Treebank release 20180919 and the ISWOC Treebank release 20160620.

Language	Number of tokens	Number of sentences	Treebank
Ancient Greek	250,455	18,173	PROIEL
Latin	225,064	19,425	PROIEL
Gothic	57,211	5,457	PROIEL
Classical Armenian	23,513	1,916	PROIEL
Old Russian	235,275	24,716	TOROT
Old Church Slavonic	58,269	6,350	PROIEL
Old Church Slavonic	82,007	8,371	TOROT
Old English	29,406	2,536	ISWOC
Old French	2,340	137	ISWOC
Old Portuguese	36,595	2,027	ISWOC
Old Spanish	54,661	2,615	ISWOC

parallel text that allows for cross-linguistic comparison of phenomena like word order, anaphoric expressions, definiteness, background events and discourse particles.

To achieve this the New Testament texts were annotated with morphosyntactic and information-structure annotation, and then aligned so that words in, for example, the *Vulgate* were linked to the words that they translate to in the Greek *New Testament*.

The PROIEL Treebank has since been expanded with other texts in Latin and Ancient Greek, which have been morphosyntactically annotated. Since the end of the original PROIEL project, the long-term objective has been to expand the treebank to the point where it contains – to the extent it is practically possible – representative samples from different periods and genres. This is why, for example, the Latin section of the treebank now includes not just the Vulgate and texts from the classical canon, like Caesar's *Gallic War*, but also works like the Late Latin *Peregrinatio Aetheriae* and sections of Palladius' agricultural handbook, and at the time of writing Petronius' *Satyricon* and samples from Plautus are being prepared.

In parallel to the continued expansion of the PROIEL Treebank, the ISWOC Treebank and the TOROT Treebank were set up. The ISWOC Treebank contributes

samples from Old English, Old French, Old Spanish and Old Portuguese, while the TOROT Treebank contributes a large and expanding selection of texts from Old Russian and Old Church Slavonic. Both are modelled on the PROIEL Treebank and were designed to be fully compatible. They therefore adhere to the same annotation scheme, were built using the same annotation process and rely on the same data representation (Eckhoff et al. 2018).

Using the same annotation scheme offers a range of advantages. For linguists using the treebank the main advantage is that it becomes possible to test cross-linguistic hypotheses, but it also significantly simplifies the process of building a treebank if resources can be combined to design shared guidelines and build shared annotation infrastructures that reflect best practices.

The Universal Dependencies project (Nivre et al. 2016) is today the largest collection of treebanks that have been harmonised in this manner, and Universal Dependencies have become the de facto standard within computational linguistics. The PROIEL Treebank predates Universal Dependencies and uses a different annotation scheme, but the PROIEL-style of annotation can be automatically converted to Universal Dependencies. The conversion relies on some heuristics but work is ongoing to align the PROIEL-style of annotation with Universal Dependencies so that these heuristics can be eliminated.

2.1 The annotation scheme and the annotation process

The PROIEL-style of annotation is based on multiple levels of annotation. Lemma, part of speech and morphological features are annotated at the morphological annotation level. The syntactic annotation level includes labelled dependencies, as well as a combination of enhanced (or 'secondary') labelled dependencies and empty elements for representing syntactic phenomena that involve gaps, coindexing or displacement. The information structure level has annotation for givenness and anaphoric reference chains. The alignment level contains links between elements that are translational equivalents in two texts. Finally, the semantic level is used for free classification of data according to criteria like aspect or lexical semantics.

Each annotation level allows for annotation of individual tokens. Some levels are also defined for larger textual units like sentences or paragraphs, but the annotation process itself is designed around sentences as the minimal unit. As annotation of a text progresses, each sentence is individually assigned an 'annotated' or 'reviewed' status, where 'annotated' indicates that the sentence has been annotated by the primary annotator and 'reviewed' indicates that it has also been approved by the secondary annotator. A sentence has to have

complete annotation on both the morphological and syntactic levels before it can be assigned the 'annotated' or 'reviewed' status, while the other levels of annotation are optional and can be added independently.

The annotation scheme used on the morphosyntactic level is broadly aligned with 'school grammar' in the sense that assumptions about morphology and syntax are not too different from what would be expected by students who have studied the language but not necessarily formal linguistics. The scheme by default also tries to adhere to linguistically informed conventions for the language and its philological traditions. For Latin, for example, lemmatising is based on the *Oxford Latin Dictionary* but has been adapted to make the relationship between headwords and parts of speech more predictable so that each lemma in the treebank has one and only one normalized headword form and one and only one part of speech.

Although no linguistic annotation is ever completely theory-independent, morphological annotation is generally uncontroversial as philologists and linguists of different persuasions generally follow the same conventions. Syntactic annotation is a different matter with wide-ranging disagreement among researchers. The syntactic annotation in PROIEL-style treebanks is based on dependency grammar. Dependency grammar is not well developed as a linguistic theory, but the PROIEL-flavour of dependency grammar has been enriched with formal devices that can handle syntactic structures like raising and control. The implementation of these devices and the specific analyses of structures with 'gaps' or long-distance dependencies is based on Lexical-Functional Grammar (Kaplan and Bresnan 1982, Bresnan 2001), whose functional structures were in turn influenced by dependency grammar. Grammatical functions like subject and object are primitives in Lexical-Functional Grammar and this assumption has also been carried over into PROIEL-style dependency grammar along with Lexical-Functional Grammar's criteria for identifying these grammatical functions.

Dependency grammar-based annotation was chosen over an annotation scheme rooted in constituency structure in part because of its near-universal adoption in current computational work, and in part because it makes it possible to annotate free-word-order languages consistently. Haug (2015) discusses the latter point in more detail, as well as broader methodological motivations and the practical implications of this choice.

The details of the syntactic annotation scheme and the precise handling of specific syntactic structures are complex and well beyond the scope of this chapter, which aims to give only a brief overview of the key characteristics of the treebanks. For further details on the morphosyntactic annotation scheme the reader is directed to the overviews by Haug and Jøhndal (2008), Haug,

Jøhndal et al. (2009), Haug, Eckhoff et al. (2009) and Eckhoff et al. (2018), while the design of the annotation scheme for information structure is described in Haug et al. (2014).

3 Long-term scalability and maintenance challenges

A number of early design choices contributed to the success of treebanks that use the PROIEL-style of annotation. Annotation requires specialist knowledge, so it is crucial to be able to recruit students and researchers across the world as annotators. This requires a tool that supports distributed annotation and that does not have to be installed on the annotator's computer, as this would have required us to provide technical support to annotators. We also needed a tool that could be tailored to the evolving annotation scheme and allow us to make continuous improvements to the software without disrupting annotators. No such tool existed in 2008 when work on the PROIEL Treebank started. We therefore opted to develop our own annotation tool as a web application.

The use of dependency grammar and the organisation into multiple levels of annotation, in which each level is independent and can be conceptualised either as a graph with nodes and edges or as pairs of tokens and feature structures, allowed for a flexible data model that could be mapped onto standard technologies for data representation and storage like XML and relational databases, and it permitted researchers to work independently, adding other annotation levels when resources and expertise became available.

Treating the sentence as the smallest unit that can be annotated and reviewed on its own is also a design decision that has worked well in practice as it made it possible to release data in batches, even when texts were not completely ready, and to preserve the history of changes in a practical way.

Finally, the Lexical-Functional Grammar-influenced variety of dependency grammar has proven to be easy for annotators with philological training to learn and apply consistently. It also allows for some flexibility in designing consistent analyses of syntactic structures across languages when there is disagreement in the linguistic literature on what the correct analysis is.

Other design choices have in hindsight proven to be suboptimal or have blocked progress. The model of having a primary annotator with a secondary annotator as a reviewer was put in place to ensure consistency while the annotation scheme was still being developed, and was subsequently used to ensure that the three treebanks were compatible and used formal devices in the same

way. This relied on extensive coordination between reviewers and centralised training of annotators. This approach worked well when several annotators were working intensively on annotating multiple texts in parallel but is not cost-effective today when only a few annotators occasionally work on expanding the treebank.

The process for developing documentation was not integrated with the annotation tool itself. Unfortunately, documentation efforts have therefore not kept up with annotation and the documentation is neither consolidated nor complete.

On the technological side, the annotation tool is monolithic, so it is hard to break it up or replace components. This makes it challenging to modify it or the data model that it uses. This is a particular issue in two areas. First, it has hampered integration with external automated taggers and parsers, which is necessary since the tool itself only has built-in support for generating suggestions using finite-state transducers or by looking up the annotation that an annotator has already chosen for a token with the same surface form. Second, it has slowed down efforts to address weaknesses in the data model, which is a particular concern as the data model lacks support for sub-token annotation, e.g. annotation of compound words or infixes.

In combination these challenges now constitute a significant barrier to further expansion of the treebanks and are risk factors when it comes to long-term maintenance and accessibility.

3.1 Syntacticus

To address the long-term scalability, maintenance and accessibility challenges, we launched Syntacticus in 2018. The aims of Syntacticus are (1) to increase the visibility, accessibility and discoverability of the PROIEL, ISWOC and TOROT treebanks, (2) to develop processes for long-term maintenance, (3) to improve the scalability of the annotation process and (4) to provide an open infrastructure platform for other researchers working on less-resourced, historical languages. These are ambitious aims that will take time to achieve. Aim 4, in particular, is a long-term aspiration. Aims 1 and 2, on the other hand, are crucial for ensuring that the treebanks remain accessible and reliable. Aim 3, in turn, is a requirement if continued expansion is going to be economically feasible.

Visibility, accessibility and discoverability (aim 1) have been addressed by setting up a dedicated website for Syntacticus (http://syntacticus.org) that provides much more direct access to data from the treebanks than before. Crucial elements include removing all registration barriers, incorporating elements of

the familiar search-engine paradigm in the user interface and making more of the treebank data indexable by search engines. We have also included direct access to data that have been synthesised from treebank data like dictionary resources that are automatically generated from the morphosyntactic annotation. The *Varangian Rus Project* (Eckhoff and Berdičevskis 2016) has in turn built an Old Russian dictionary with glosses in Russian and English on top of the synthesised dictionary for Old Russian.

At the time of writing much work remains to be done before the Syntacticus site is mature and satisfies our requirements, but the process for achieving this is well understood and achievable given recent advances in web technology and the broad availability of suitable open-source software components. The remainder of this paper is devoted to discussing how we aim to address annotation scalability (aim 3), which presents significant challenges for low-resourced languages.

4 Scaling morphosyntactic annotation

Manual annotation of lemma, part of speech and morphological features is time-consuming, error-prone and very tedious for annotators. The practical experience from PROIEL, ISWOC and TOROT has shown that annotation speed increases and the error rate decreases when annotators are provided with some automated assistance, such as pre-populated annotation fields that they can correct or a list of suggested annotations that they can choose from. The effect is positive even when this assistance is very crude and generated using simple methods, such as looking up annotations that have already been made earlier in the text, ranking them by frequency and serving them to annotators as suggestions.

More sophisticated and higher-accuracy assistance can be provided if we use automated taggers, parsers and other techniques in natural-language processing (NLP). The difficulty here is that historical languages are, in NLP jargon, low-resource languages. This means that the data sets and models that are prerequisites for applying many NLP techniques do not usually exist and have to be built largely from scratch. For example, in order to use a statistical part-of-speech tagger you would have to train the tagger using a corpus that has already been annotated with parts of speech.

While some required language resources, like part-of-speech-tagged corpora, do exist for the most widely studied historical languages, they may not be suitable for the task. It is common for such resources to be too small, or to suffer