Proceedings of the CLTW 4 @ LREC2022 , pages 85–93
Marseille, 20-25 June 2022

Towards Coreference Resolution for Early Irish

Mark Darling[1], **Marieke Meelen**[2], **David Willis**[1]
[1]University of Oxford, [2]University of Cambridge
{mark.darling,david.willis}@ling-phil.ox.ac.uk, mm986@cam.ac.uk

Abstract

In this article, we present an outline of some of the issues involved in developing a semi-supervised procedure for coreference resolution for early Irish as part of a wider enterprise to create a parsed corpus of historical Irish with enriched annotation for information structure and anaphoric coreference. We outline the ways in which existing resources, notably the POMIC historical Irish corpus and the Cesax annotation algorithm, have had to be adapted, the first to provide suitable input for coreference resolution, the second to cope with specific aspects of early Irish grammar. We also outline features of a part-of-speech tagger that we have developed for early Irish as part of the first task and with a view to expanding the size of the future corpus.

Keywords: Old Irish, Middle Irish, Information Structure, Coreference Resolution, Low-Resource NLP

1. Introduction

Because of their unique position, having both lexical and functional characteristics, pronouns form an excellent starting point for both diachronic as well as crosslinguistic research as they are widely assumed to proceed through a cycle of reduction from independent pronoun to inflectional affix and zero elements or 'null pronouns' (Siewierska, 1999; Van Gelderen, 2011). Links of pronouns to their referents are established through either linguistic, contextual licensing or extra-linguistic factors related to discourse. Much of the literature on pronouns, however, is either grammar-oriented, focusing on the correlations of, for example, null pronouns with other parts of the grammar ('rich agreement', a rich determiner system, or word order). Other authors focus solely on the information structure (IS) of anaphor–antecedent relations and contextual licensing. A crucial question that needs to be answered, however, is if and how these morphosyntactic and information-structure dimensions interact. In order to investigate how the presence or absence of subject pronouns reflects the flow of new and old information and of changing topics of discourse, we need a deeply annotated corpus, enriched with morphosyntactic and information-structural annotation. In this article, we report on how such a corpus can be developed for early Irish using rich annotation and semi-supervised coreference resolution.

Coreference resolution is an NLP task developed in the 1960s that involves determining all referring expressions that point to the same real-world entity. A referring expression in this case is often either a noun phrase (NP) (*the woman*, *Mary*) or a pronoun (*she*), either of which refer to an entity in the real world known as the referent (a specific woman evident in the context) (Sapena et al., 2013). The goal of a coreference-resolution system is to output all the coreference chains of a given text, thus identifying *a woman*, *Mary* and *she* as coreferring in the sequence *A woman walked in. It was Mary. She started to speak.*. This may allow us to gain insights into not only pronominal forms and functions, but also into topic chains and shifts (if the text continues *When she had finished, John asked a question*, then the topic has shifted from Mary to John). Irish is particularly interesting within this context, since its use of subject pronouns has changed considerably over time, it having been essentially a null-subject language in the earliest documentation, and gradually developing a requirement for overt subject pronouns in most parts of the verbal paradigm.

In this article, we focus on developing semi-automatic coreference resolution for Old Irish. We start by evaluating existing language resources for early Irish and assessing how these need to be extended to be suitable for our task (Section 2). In Section 3, we outline the necessary preprocessing stages as well as presenting an automatic part-of-speech (POS) tagger for Old Irish, before turning to our main task of coreference resolution in Section 4.

2. Current Irish Corpora

One aspect of the workflow is the building of a diachronic corpus of Irish, annotated with part of speech and information-structural features. Existing Irish corpora can be divided into two categories. First, there are large online text corpora, with minimal annotation. These include:

- the Thesaurus Linguae Hibernicae (TLH) (Kelly and Fogarty, 2006);
- the Corpus of Electronic Texts (CELT);
- the Historical Irish Corpus.

Where these are annotated at all, this annotation is generally limited to standard Text Encoding Initiative (TEI) annotation for text structure, and does not extend to POS tagging or annotation of syntactic features. Second, there are a few linguistically annotated corpora. Examples of this type of corpus include:

- Parsed Old and Middle Irish Corpus (POMIC) (Lash, 2014), a selection of fourteen short texts, largely from the Old Irish period, manually annotated with POS tags and constituent structure;
- the Universal Dependency (UD) treebanks of Old Irish and Middle Irish, which are currently works in progress, and are not included in version 2.10 of the UD treebanks. The Old Irish treebanks currently available consist of the St Gall glosses on Priscian (around 22,000 tokens), while for Middle Irish there are around 800 tokens of *Scéla Mucce Meic Dathó* (The Tale of Mac Dathó's Pig), not all of which have been tagged;
- The online database of the St Gall glosses, on which the UD Old Irish treebanks of the same text are based;
- The Corpus Palaeohibernicum, which contains over 70 annotated Old and Middle Irish texts, in spreadsheet form.

Clearly, none of the large online corpora are sufficient as they stand for research into the diachrony of subject pronouns in Irish, but they do provide a valuable resource of digitised texts. Of the existing annotated corpora, POMIC is the most immediately useful for our purposes, as it consists of Penn-style tagged and parsed texts. It lacks IS annotation, however, which is required for our study of how the use of subject pronouns changes over time in Irish. We are therefore building a larger, POS-tagged corpus, which will be augmented with IS annotation. The other linguistically annotated resources detailed above may, however, prove useful as training data for a POS tagger, and as future target texts for incorporation into the corpus. This corpus is being built to conform to the standards of the ongoing Parsed Historical Corpus of the Welsh Language (PARSHCWL) (Meelen and Willis, 2021; Meelen and Willis, 2022), a Penn-style treebank of historical Welsh (Willis and Mittendorf, 2004b) based on the Historical Corpus of the Welsh Language 1500–1850 (HCWL) (Willis and Mittendorf, 2004a).

2.1. POMIC

POMIC consists of fourteen manually annotated texts with a Penn-style tagset adapted for Old and Middle Irish. The annotation scheme was adapted from the 2010 version of the manual for the Penn Corpora of Historical English (Santorini, 2022). The texts span the period between around 700 and 1100 CE. We use POMIC as a starting point. The majority of the texts – ten of fourteen – are at least arguably of Old Irish date, meaning that they most likely predate the 10th century CE, generally taken as when Old Irish gives way to Middle Irish (McCone, 1996, p. 140). In practice, distinguishing Old from Middle Irish is not simple, but the preponderance of Old Irish material in the POMIC data means that it can be used to train a reasonably accurate tagger for Old Irish.

2.2. Necessary Extensions

Although useful as a starting point, POMIC requires a number of extensions for our purposes. In the first place, the manual tagging process understandably led to some errors, which need correcting in order to use POMIC as a training corpus for a POS tagger. For example, the tag and token of the perfective particle *ro*, normally `(RO ro)` in POMIC, are occasionally inverted, giving `(ro RO)`. The POS tagger is case-sensitive, so this will be interpreted as a separate token and tagset, reducing the overall accuracy. Furthermore, the annotation scheme, and particularly the use of compound tags, leads to a very large number of discrete tags, significantly complicating the process of training a tagger. We therefore reduce the number of tags by either splitting the compound tags into individual tokens or by reducing them to a single tag, detailed further below. We also remove discrete tags for initial-consonant mutations (tagged by Lash as `NAS`, `LEN`, and `GEM`).

We also need to add information not included in the POMIC annotation scheme. The existing tagged texts lack the following information which could be salient for the research questions we want to answer:

- person–number information for verbal forms: this information will be useful for investigating whether there are any patterns in the use of pronouns that correlate to specific persons and numbers of subjects;
- individual tokens for infixed and suffixed pronouns – these are particularly important, as they can be involved in coreference chains, and can refer to separate entities from the verb with which they form a single prosodic word;
- person–number information for pronominal forms and conjugated prepositions, which will be useful for establishing coreference further downstream.

3. Preprocessing and POS tagging

Creating a POS tagger or any other dedicated NLP tool for a historical language is challenging for a number of reasons. First and foremost, there are issues of data scarcity: historical languages are often classified as

extremely low-resource from an NLP point of view,[1] because the amount of data is necessarily finite due to the surviving attestation, and often also limited in range. In addition, not all data is easily available or accessible. Finally, if material is available, it often requires much preprocessing, because orthography is not standardised.

The situation is further complicated by the fact that historical languages are not only low-resource but also under-researched from an NLP point of view: whereas there are numerous off-the-shelf tools available for basic preprocessing and annotation in modern varieties (even modern varieties of Irish and other Celtic languages), this is not the case for their historical counterparts. Since Old Irish differs significantly from Present-Day Irish, we cannot simply apply or even easily modify existing tools, e.g. tokenisers, morphological transducers and POS taggers (Uí Dhonnchadha, 2002; Uí Dhonnchadha et al., 2003; Uí Dhonnchadha and Van Genabith, 2006).[2]

The lack of NLP resources for early Irish ultimately reflects the fact that the extremely complex inflectional system, the phonological challenges of mutated initial consonants, and the orthographic inconsistencies, even of edited texts, significantly complicate the processing of early Irish source material. There has been some work on producing a general POS tagger for early Irish (Lynn, 2012), but this was, by the author's own admission, "rudimentary": the results published show that the tagger could only differentiate between types of part of speech (verb, noun, etc.), but no finer detail of inflection could be distinguished. More recently, there has been work to develop computational methods for identifying and tagging Old Irish weak verbs, building on Uí Dhonnchadha's work on Modern Irish (Fransen, 2019; Fransen, 2020b; Fransen, 2020a). While this work deals with the right period in the history of Irish for our work, we require a tagger that functions more comprehensively, meaning that we cannot make use of Fransen's previous work in this area.

Efforts have been made in recent years to develop an Old Irish lemmatiser (Dereza, 2016; Dereza, 2018; Dereza, 2019), trained on the Dictionary of the Irish Language, but even the most recent version cannot lemmatise everything (accuracy ranges from 64.9% for unknown tokens to 99.2% for known tokens) and it was tested on a rather small corpus ($83k$ tokens). We use this for new texts as, despite the error rate, it is still an improvement on the complete absence of lemmatisation. It does not, however, address normalisation of orthography, which is why we deal with this separately, both for POMIC, used as a starting point, as well as for new texts.

In the following subsections we discuss all stages of preprocessing and POS tagging, which are necessary prerequisites to successful performance of coreference Resolution.

3.1. Normalisation

Even in POMIC, which is based on published text editions, there is orthographic variation. Some of this is an unavoidable consequence of working with historical data, from a period prior to standardisation. Additionally, the texts in POMIC were edited by various editors, following different editorial practices: some editions are more diplomatic, more or less directly reflecting the manuscript, while others attempt to restore a reconstructed "original" text by undoing modernisations or errors of later scribes.

One type of variation that can be controlled relatively easily at an early stage is the spelling of long vowels, which in POMIC are indicated either with macrons (ā, ē, ī, ō, ū) or with an acute accent (á, é, í, ó, ú). The two spelling practices are both used in editions of early Irish texts to denote long vowels, the former when a long vowel is not marked in a manuscript, the latter for when it is indicated with a diacritic. For the purpose of training a tagger on a small training corpus,[3] it is preferable to have just one spelling for each long vowel in the language as it reduces the number of unique tokens. Thus, for the moment at least, we automatically replace the spellings with macrons with those with acute accents. However, as the corpus develops and the accuracy of the tagger improves, we will be able to reintroduce spelling variation, reducing the amount of normalisation required during preprocessing. In the first instance, we expect this to reduce the accuracy of the tagger, but, with enough tokens in the corpus, it should be possible to retain a reasonable level of accuracy with a greater degree of orthographic variation.

3.2. Splitting and Combining Tokens

In the POMIC annotation scheme, an entire verbal complex (a prosodic element that can consist of preverbs, infixed or suffixed pronouns, aspectual particles, and the finite verb) is treated as a single token. In order to be able to use the POMIC texts for coreference resolution, these must be split into individual tokens. Consider:

[1] Regarding early Irish specifically, note the reference to it as an "under-resourced language" by Dereza (2019).

[2] For some historical languages, this situation has recently improved with the release of the Classical Language Toolkit (Johnson et al., 2021), but historical Irish is not presently covered by this toolkit.

[3] On the problems of orthographical variation in NLP, and the benefits and difficulties of "canonicalisation" as a way to address this issue, see Piotrowski (2012, ch. 3, 6).

(1) *do- s- raithminestar*
PV PRO-3PL call.ASP-VBD-3SG
'has called them to mind'

POMIC tags this as `(PV+X+VBD-RO)`, treating the entire verbal complex as a single token. Given that the infixed pronoun *-s-* can participate in coreference relations with other noun phrases in the text, subsuming it into a single token with the verb is undesirable. Moreover, such long tags with many variable components make it more difficult to automate the POS-tagging process with machine learning. We have to break up composite tags such as this into their constituent parts, the break point being denoted with the symbol `#`, resulting in this example being tagged as:

(2)
```
(PV do#)
(NP-OBJ (PROI-3PL s#))
(ASP-VBD-3SG raithminestar)
```

This allows us to enrich the annotation of the texts further downstream in the workflow, and should accelerate annotation of new texts.

Similarly, for a number of combinations, POMIC treats the sequence of preposition and possessive pronoun, which can form a single prosodic word in Irish, as a single token:

(3) *atá ocom chungid*
be-3SG at-my seeking-D
'she is seeking me'

POMIC tags *ocom* here as `(P+PRO$)`. We instead separate the possessive pronoun from the preposition, yielding:

(4)
```
(BEPI-3SG atá)
(PP (P oco#)
(PRO-G-1SG m)
(NP (VBN-D chungid)))
```

This means that only conjugated prepositions, which are not easily reducible to their constituent elements, are treated as single tokens (analogous to inflected verbs), while other combinations of preposition and personal pronoun are separated into discrete tokens.

There are also instances in which it is useful to combine tokens treated by POMIC as separate. This is particularly the case in stereotyped adverbial phrases, such as:

(5) *iar na bárach*
after POSSESSIVE morrow/milking.time
'the next day, tomorrow'

A particular problem presented by this collocation is that it is difficult to determine the gender of the possessive pronoun *a* (here nasalised as *na*), which anyway does not have an antecedent. In POMIC, this is treated as a prepositional phrase:

(6)
```
(PP (P ar)
(NAS n)
(NP (PRO$ a)
(N-D bárach)))
```

Given that this phrase functions as an adverb from an early stage of the language, we instead combine the tokens and tag as follows:

(7)
```
(ADV ar!na!bárach)
```

3.3. Refining the POS tagset

As well as using compound verbal tags, POMIC follows the Penn annotation scheme in including a number of compound nominal tags. These too are simplified; thus, `(ADJ+NS-G óc-ban)` 'young woman' is reduced to `(NS-G óc-ban)`. We also combine POMIC's mutation tokens with the following token, in order to avoid the corpus containing surplus tokens that might be susceptible to confusion with others that are more salient for our research questions. Representation of mutations in early Irish sources is a difficult topic in its own right, and indeed it is sometimes unclear whether a mutation should be considered a feature of the mutating or the mutated word. In our corpus, we attempt to achieve a reasonable degree of uniformity in their representation, while maintaining an awareness that this might not always be possible. Thus, in the revised corpus, (8) becomes (9).

(8)
```
(CP-ADV (C co)
(NAS m)
(IP-SUB (BED buí)))
```

(9)
```
(CP-ADV (C co)
(IP-SUB (BED-3SG mbuí)))
```

As the above examples make clear, we also enrich the POMIC tagset with person–number (and, where relevant, gender) information. This applies to verbs, pronouns, and conjugated prepositions. These alterations bring the revised corpus into alignment with the Welsh PARSHCWL corpus, and provide additional information useful for our research questions. Overall, we reduce the overall number of distinct tags to around 340, while also enriching the information contained in the individual tags.

3.4. Training a POS Tagger

POMIC gives us $30k$ tokens (including punctuation) that can be used to start training a POS tagger. This is too little material to train any off-the-shelf neural-network-based tagger, but it is enough to start incrementally training a Memory-Based Tagger such as the TiMBL MBT (Daelemans et al., 2003). Even though this is not a recently developed tool, it is one of the most effective methods for developing a POS tagger from scratch, since it can learn from such specific features as initial and final characters as well as the con-

text, yielding high rates of accuracy even for extremely small data sets (Meelen et al., 2021). To train the POS tagger, we deleted all null elements, since they will not be present in the new texts planned for the future corpus. Initial results are given below with parameter settings that are manually optimised for this specific corpus. The Memory-Based Tagger (MBT) allows for optimisation of parameters for both preceding and following context, but also for up to the first three and last three characters of the word, which is useful for morphologically rich languages with various inflectional suffixes like Old Irish; for a full list of parameter options, see Daelemans et al. (2003) .

(10) Parameters:
```
-p dwdwfWaw
-P psssdwdwdwFawaw
-M 1100 -n 5 -% 8 -O+vS
-FColumns -G
K: -a0 -k1
U: -a0 -mM -k17 -dIL
```

We do a 10-fold cross-validation to evaluate the results, measuring the global accuracy, which averages the harmonic means of all 340 unique POS tags for seen and unseen (i.e. known and unknown) tokens. For the 10-fold cross-validation, we separate a 10% test set from 90% training data to make sure we do not evaluate on training data. In order to control for variation and repetition at any point in our training data, we repeat this test-training division 10 times and evaluate the results of each round, using precision, recall and f-scores to calculate the final global accuracy:

(11) Global Accuracy: 0.751
Global Accuracy seen words: 0.829
Global Accuracy unseen words: 0.580

These preliminary results are not optimal, but they form a first step to providing new Old Irish texts with highly detailed morphosyntactic tags. Once new texts are tagged and manually corrected, they will be added to the training corpus, which will at this stage – where we have only a $30k$-token Gold Standard, but over 340 unique POS tags – improve the results significantly. In addition, when more texts are preprocessed and added to the corpus, we can create word embeddings which will allow us to test neural-network based taggers like TARGER (Chernodub et al., 2019). Improving POS tagging results is important when new texts are added to our treebank, but we leave this for future research since these results are sufficient for our main Coreference Resolution trials at hand.

4. Coreference Resolution

We use the Cesax coreference resolution algorithm (Komen, 2013) as a starting point for our Old Irish Coreference Resolution (Komen, 2019). This software was originally designed for use on historical English data, but has since been extended to include support for several other languages, including Dutch, Chechen, and Welsh. Although Irish is not yet one of the languages supported by the software in its unmodified state, some relatively simple adjustments can be made to the software's settings in order to accommodate historical Irish data. Cesax is particularly appropriate for our corpus due to the fact that it can import Penn-style treebank files for IS annotation, which can then be exported back to PSD (phrase-structure description) format, as well as to a variety of other formats, such as Folia XML.

4.1. Semi-Supervised Method

The Cesax coreference-resolution algorithm uses a set of hierarchically ordered constraints to evaluate possible solutions. It evaluates every noun phrase in the input text individually, trying to find connections and ultimately the best antecedent based on the following information:

- NP type
- grammatical role (function)
- person, gender and number

NP types include pronouns, definite/indefinite NPs, demonstratives, proper nouns, etc. Grammatical roles include subject and object (of verbs and/or prepositions) as well as possessive/genitive. Pronominal elements manifest person, gender and number in Old Irish. Non-pronominal NPs are all considered to be third-person.

In order to process and annotate Irish texts in Cesax, we have to carry out a series of tasks:

1. Define the nodes that can be involved in coreference in Irish. By default, Cesax only targets NPs, nominal *wh*-phrases, pronouns and proper nouns. This works for historical English, but misses some nodes that we want to target for coreference in Irish, meaning that the resulting coreference chains would be incomplete. We therefore edited the settings of Cesax to add conjugated prepositions and finite verbs to the possible targets for coreference. Targeting finite verbs is particularly important, since this allows us to capture null subjects in coreference chains.

2. Replace the historical English pronouns in the Cesax settings with those for Irish. At this stage, we must try to avoid including homophonous pronouns in more than one category. For example, the emphatic pronoun *som*, which can refer either to a third-person singular masculine or neuter referent, or to a third-person plural one, has to be treated as generically third-person.

3. Import texts into Cesax. Cesax converts Penn-style .psd files into XML documents, which are then saved as .psdx files. At this stage, we can also check for any pronominals or demonstratives that fall outside our existing lists of such forms (Tools > Features > Renew features of... > NP – all noun-phrase features), and add any new forms, shown in the "Errors" tab, to the relevant categories.

4. Perform a manual check for conflicts by opening the .psdx file in an XML editor. Due to the use of wildcards to capture all of the possible forms of pronouns in our texts, some forms are assigned more than one classification. For example, the 1sg. emphatic pronoun *sa* is sometimes misclassified as "unknown" by the software. This is due to the presence of the string "s?" in the category "Pers", used to capture the Class A infixed pronoun *-s-* (tokenised in our corpus as `PROI-3SGF s#` or `PROI-3PL s#`). Performing a check for "unknown;" or any other conflicts in the person–gender–number (PGN) features of the NPs in the XML document, and correcting them there, avoids problems when running semi-automatic coreference resolution.

5. Run semi-automatic coreference resolution on the text. Cesax looks for likely coreference targets by assessing the text against a series of constraints, in order to suggest what the most likely coreference for a given NP might be.

4.2. Targets

Several part-of-speech types can act as target for coreference in our Irish texts. These include:

- pronouns
- NPs
- inflected verbs and prepositions
- emphasising particles (*notae augentes*)

Some of these are not automatically targeted by Cesax. Cesax supports targeting pronouns and NPs, as these are also potential targets for coreference in English. Emphasising particles (*notae augentes* in some scholarship) are tagged as pronouns in our text files, hence can be easily targeted for coreference. Inflected verbs and prepositions must be added manually to the categories to be targeted, however. This is done by adding the terms P-*, VB*-[1234]*, COP*-[1234]*, and BE*-[1234]* to the tab "Phrase Types" in the Cesax settings.

4.3. Constraints

The coreference-resolution algorithm in Cesax tests NPs (and, with our modifications, inflected verbs and prepositions) against various constraints in order to establish the most likely coreferent for a given NP. For now, these constraints are being retained, but will be refined if it is found that any of them do not apply to Irish as well as they do to historical English. The algorithm assigns each possible coreferent a score based on how many of the constraints it violates; the higher the score, the less likely a candidate is considered as a target for coreference. For example, the further a coreference source is from its potential target (for example, the further a pronoun is from an NP it might refer to), the less likely it is deemed to be that there should be a direct coreferential link between the two, and the algorithm will instead attempt to identify a target nearer to the source. The constraints are tested in a given order, which the Cesax manual itself notes is designed for Modern British English. It may therefore require additional adjustment and refinement for Irish, but it nevertheless provides a good starting point.

5. Case Studies

In the following case studies, we demonstrate how Cesax can be used to conduct coreference resolution on Irish texts, and how the result can subsequently be exported to other formats for future analysis. At present, the accuracy rate of the semi-automatic coreference resolution is low: on a test passage of four sentences, the algorithm selected the correct antecedent in just under 14% of cases. This is initially a disappointing result, but there are some positive trends that can be identified in the links that the algorithm makes correctly. The results become more accurate the further into a passage of text the algorithm is allowed to run. This is to be expected, given that the first section of any given passage of text is likely to include very few coreferential links, whereas later sections of the text are likely to contain pronouns (or finite verbs) that refer to NPs introduced in the earlier sections. Furthermore, the algorithm regularly correctly identifies the link between a finite verb with null subject and its nominal antecedent in a previous clause. Since this is a type of coreference target we added specifically for early Irish, this is an encouraging result. We are continuing to work to improve the results of the semi-automatic process through refining our settings.

5.1. NP

Fig. 1 shows a coreference chain for the proper noun (personal name) *Laisrén*. This coreference chain was generated semi-automatically using Cesax's coreference algorithm, and corrected manually. The chain for *Laisrén* includes the proper noun *Laisrén* itself; the finite verbs *áin* '(he) fasted', *cúala* '(he) heard' (twice), *glúais* '(he) moves', *to-ocaib* '(he) raises', *do-beir* '(he) bears, makes', and *con-aca* '(he) saw' etc.; the possessive pronoun *na* and *a* 'his'

(three times); the infixed pronoun *-n-* 'him' and the conjugated prepositions *fair* 'over him' and *fris* 'to him'. It also crosses other coreference chains, such as that between *in guth* 'the voice' and *sodain* 'at that [voice]'. Cesax makes an error when processing this semi-automatically: due to *clúana* 'of Clúain' in the first line being a third-person singular NP, the algorithm automatically assumes that it is coreferent with *Laisrén*. This is corrected manually by deleting the coreference link.

5.2. Emphasising Particles Case Study

Language-specific adaptations of the Cesax algorithm are likely to improve its performance. One area that seems like a plausible area for improvement concerns the emphasising particles (*notae augentes*). Subject pronouns are obligatorily null with Old Irish finite verbs. However, in some contexts, verbs appear with emphasising particles. These particles have sometimes been analysed as pronouns, and this is how they are tagged in the corpus. It has also been suggested that there is an interaction between use of these elements and the marking of topics (Griffith, 2008; Griffith, 2011).

Correct coreference relations for an example containing multiple emphasising particles are shown in Fig. 2 (second person singular) and 3 (first person singular). It seems that use of the particles indicates repeated alternation between first and second person as the discourse topic. In this case, the existing Cesax algorithm produced the correct coreference resolution, since the two coreference chains clearly differ in person. Where they do not, the coreference-resolution algorithm could perhaps be improved by the addition of a resolution rule to disfavour an immediately preceding element as the antecedent for an emphasising particle.

6. Postprocessing

Once coreference resolution has been determined and corrected, the result is re-exported to PSD format, as well as other formats such as Folia XML, through Cesax. In PSD format, the coreference annotation is expressed as features added to the node of the original token. The following example demonstrates a second-person singular emphasising particle *su*, annotated as representing a subject grammatical function with information-structure marked as identical to a preceding element in the coreference chain.

(12)
```
(NP-SBJ (FS-IPdist 0)
(FS-RefType Identity)
(FS-NdDist 1)
(FS-GrRole Subject)
(FS-PGN 2s)
(FS-NPtype Pro)
(PRO-2SG (FS-IPdist 0)
(FS-RefType Identity)
(FS-NdDist 2)
(FS-GrRole unknown)
(FS-PGN 2s)
(FS-NPtype Pro)
(LEX su)))
```

7. Conclusion

In this article, we have considered the ways in which it is necessary to adapt existing resources to develop a historical parsed Irish treebank with rich mark-up information structure and coreference. To fully utilise the POMIC corpus for our needs, preprocessing was necessary, notably separation of compound tags into individual tokens, making those tokens accessible to the coreference-resolution algorithm in Cesax. For early Irish, it was necessary to adapt the Cesax algorithm so that finite verbs and conjugated prepositions can be incorporated into coreference chains. Further refinement of the process of semi-automatic coreference resolution may be motivated by other specific aspects of early Irish grammar such as the emphasising particles. Other such refinements, both to the POS-tagging and coreference-resolution algorithm, may be required as more texts are added to the corpus.

8. Acknowledgements

We gratefully acknowledge the support of the Arts and Humanities Research Council via the award of AHRC–DFG UK–German collaborative research project in the humanities, AHRC Research Grant AH/V00347X/1 for the project 'The history of pronominal subjects in the languages of northern Europe'.

9. Bibliographical References

Daelemans, W., Zavrel, J., van den Bosch, A., and Van der Sloot, K. (2003). MBT: Memory-based tagger. *Reference Guide: ILK Technical Report-ILK*, pages 3–13.

Dereza, O. (2016). Building a dictionary-based lemmatizer for Old Irish. In *Actes de la conférence conjointe JEP-TALN-RECITAL*, volume 6, pages 12–17.

Dereza, O. (2018). Lemmatization for ancient languages: Rules or neural networks? In Dmitry Ustalov, et al., editors, *Artificial Intelligence and Natural Language*, pages 35–47. Springer.

Dereza, O. (2019). Lemmatisation for under-resourced languages with sequence-to-sequence learning: A case of early Irish. In *Proceedings of Third Workshop "Computational linguistics and language science"*, volume 4, pages 113–124.

Fransen, T. (2019). *Past, present and future: Computational approaches to mapping historical Irish cognate verb forms*. PhD, School of Linguistic, Speech and Communication Sciences, Trinity College Dublin.

Fransen, T. (2020a). Automatic morphological analysis and interlinking of historical Irish cognate verb forms. In Elliott Lash, et al., editors, *Morphosyntactic Variation in Medieval Celtic Languages: Corpus-based approaches*, pages 49–84. De Gruyter Mouton.

Fransen, T. (2020b). Automatic morphological parsing of Old Irish verbs using finite-state transducers. *LanguageLeeds Working Papers*, 1:15–28.

Griffith, A. (2008). The animacy hierarchy and the distribution of the *notae augentes* in Old Irish. *Ériu*, 58:55–75.

Griffith, A. (2011). Old Irish pronouns: Agreement affixes vs. clitic arguments. In Andrew Carnie, editor, *Formal approaches to Celtic linguistics*, pages 65–94. Cambridge Scholars Publishing, Newcastle upon Tyne.

Johnson, K. P., Burns, P. J., Stewart, J., Cook, T., Besnier, C., and Mattingly, W. J. B. (2021). The Classical Language Toolkit: An NLP framework for pre-modern languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 20–29, Online, August. Association for Computational Linguistics.

Komen, E. R. (2013). Predicting referential states using enriched texts. In *The Third Workshop on Annotation of Corpora for Research in the Humanities (ACRH-3)*, page 49.

Lynn, T. (2012). Medieval Irish and computational linguistics. *Australian Celtic Journal*, 10:13–28.

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

Meelen, M. and Willis, D. (2021). Towards a historical treebank of Middle and Early Modern Welsh, part I: Workflow and POS tagging. *Journal of Celtic Linguistics*, 22:125–154.

Meelen, M. and Willis, D. (2022). Towards a historical treebank of Middle and Modern Welsh: Syntactic parsing. *Journal of Historical Syntax*, forthcoming.

Meelen, M., Roux, E., and Hill, N. (2021). Optimisation of the largest annotated Tibetan corpus combining rule-based, memory-based & deep-learning methods. *Transactions on Asian and Low-Resource Language Information Processing*.

Piotrowski, M. (2012). Natural language processing for historical texts. *Synthesis lectures on human language technologies*, 5(2):1–157.

Santorini, B. (2022). Annotation manual for the Penn Parsed Corpora of Historical English and the Parsed Corpus of Early English Correspondence.

Sapena, E., Padró, L., and Turmo, J. (2013). A constraint-based hypergraph partitioning approach to coreference resolution. *Computational Linguistics*, 39(4):847–884.

Siewierska, A. (1999). From anaphoric pronoun to agreement marker: Why objects don't make it. *Folia Linguistica*, 33:225–251.

Uí Dhonnchadha, E. and Van Genabith, J. (2006). A part-of-speech tagger for Irish using finite-state morphology and constraint grammar disambiguation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 2241–2244, Genoa, Italy, May. European Language Resources Association (ELRA).

Uí Dhonnchadha, E., Pháidín, C. N., and Genabith, J. V. (2003). Design, implementation and evaluation of an inflectional morphology finite state transducer for Irish. *Machine Translation*, 18(3):173–193.

Uí Dhonnchadha, E. (2002). A two-level morphological analyser and generator for Irish using finite-state transducers. In *LREC*.

Van Gelderen, E. (2011). *The linguistic cycle: Language change and the language faculty*. Oxford University Press, Oxford.

Willis, D. and Mittendorf, I. (2004b). Ein historisches Korpus der kymrischen Sprache. In Erich Poppe, editor, *Keltologie heute: Themen und Fragestellungen*, pages 135–42. Nodus, Münster.

10. Language Resource References

[Royal Irish Academy]. (no date). *Historical Irish Corpus 1600–1926*. Royal Irish Academy.

Andersen, Erik. (no date). *Universal Dependencies treebank of Middle Irish*.

Bernhard Bauer and Rijcklof Hofman and Pádraic Moran. (2017). *St Gall Priscian Glosses*.

Chernodub, A., Oliynyk, O., Heidenreich, P., Bondarenko, A., Hagen, M., Biemann, C., and Panchenko, A. (2019). TARGER: Neural argument mining at your fingertips. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 195–200.

Doyle, Adrian. (no date). *Universal Dependencies treebank for the Old Irish glosses of St. Gall*.

Kelly, Patricia, Brady, Niall and Fogarty, Hugh. (2006). *Thesaurus Linguae Hibernicae*. School of Irish, Celtic Studies, Irish Folklore & Linguistics, University College Dublin.

Komen, Erwin. (2019). *CESAX: Coreference Editor for Syntactically Annotated XML corpora*. Radboud University Nijmegen.

Lash, Elliott. (2014). *POMIC: The parsed Old and Middle Irish corpus. Version 0.1*. Dublin Institute for Advanced Studies.

Stifter, David, Bauer, Bernhard, Qiu, Fangzhe, Lash, Elliott, White, Nora, Nguyen, Truc Ha, Felici, Francesco, Osarobo, Godstime, Ji, Tianbo, Ganly, Ellen, Nooij, Lars, and Bulatovas, Romanas. (2020). *Corpus Palaeohibernicum*.

Willis, David and Mittendorf, Ingo. (2004a). *Historical Corpus of the Welsh Language 1500–1850*. University of Cambridge.

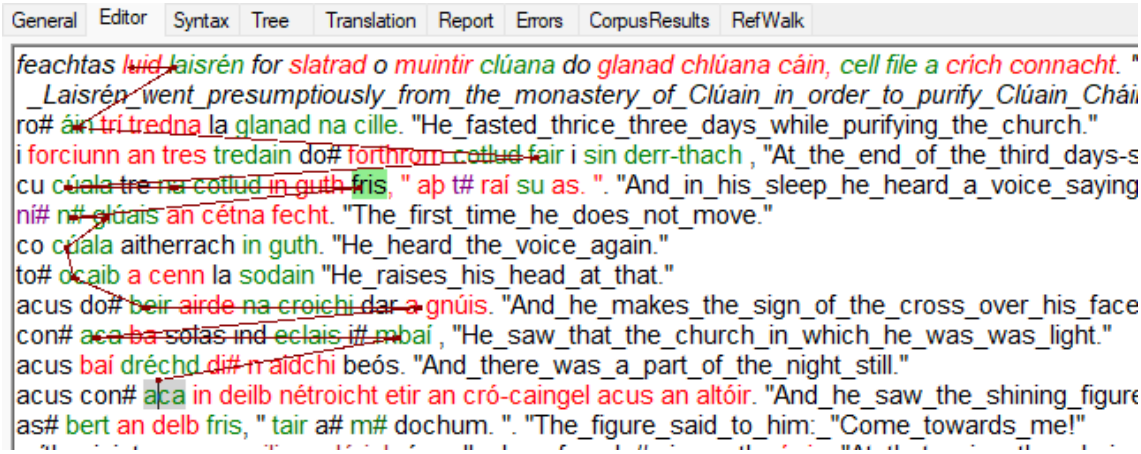

Figure 1: Coreference chain for *Laisrén*

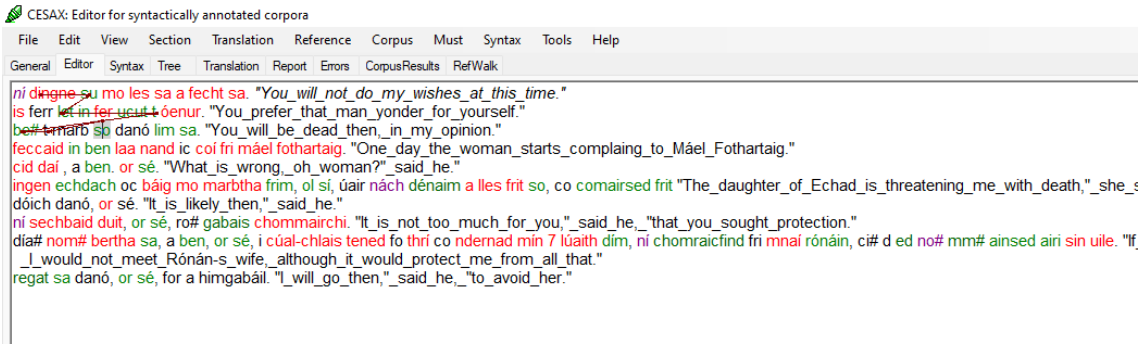

Figure 2: Coreference chain for 2sg *nota augens*

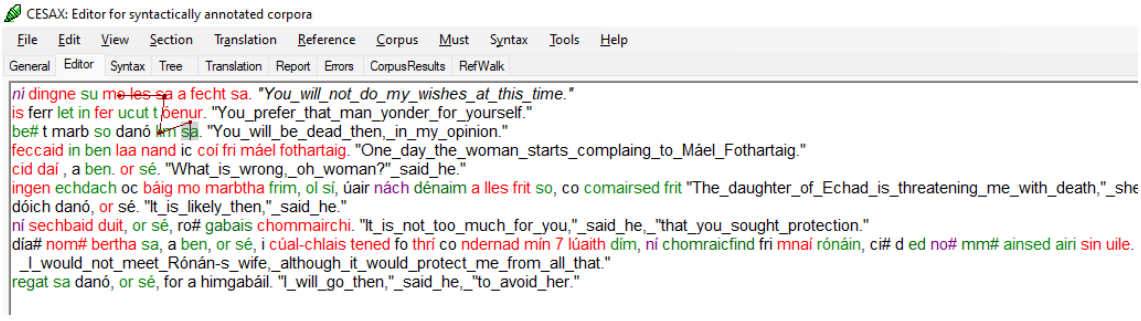

Figure 3: Coreference chain for 1sg *nota augens*