Towards a historical treebank of Middle and Early Modern Welsh
Part I: Workflow and POS tagging

Abstract

This article introduces the working methods of the Parsed Historical Corpus of the Welsh Language (PARSHCWL). The corpus is designed to provide researchers with a tool for automatic exhaustive extraction of instances of grammatical structures from Middle and Modern Welsh texts in a way comparable to similar tools that already exist for various European languages. The major features of the corpus are outlined, along with the overall architecture of the workflow needed for a team of researchers to produce it. In this paper, the two first stages of the process, namely pre-processing of texts and automated part-of-speech (POS) tagging are discussed in some detail, focusing in particular on major issues involved in defining word boundaries and in defining a robust and useful tagset.

1 Introduction

The last thirty years have witnessed the development of more and more sophisticated textual corpora for the study of large-scale linguistic history.[1] Early historical corpora consisted of collections of texts or samples of texts, often balanced to ensure even coverage over time and by text type, and permitted only searches that could be linked to a particular string of characters. The Helsinki Corpus of English Texts, published in 1991 (Kytö 1996), was a pioneering corpus of this type. In syntactic research, this essentially limited the researcher to looking at phenomena, such as the syntax of auxiliaries, for which data could be extracted by searching for a given set of words and their spelling variants. As time has gone on, the creation of parsed corpora and historical treebanks has paved the way for large-scale searching of historical texts to examine variation and change in more abstract syntactic structures such as word order. For a number of languages, there are now well-established historical corpora that have been fully annotated for both morpho-syntactic category (‘part of speech’) and syntactic structure. A series of corpora based around the various Penn parsed corpora of historical English, such as the Penn–Helsinki Parsed Corpus of Middle English (Kroch and Taylor 2000), have adopted a broadly common framework, providing a parallel format for part-of-speech information, and adopting a constituency-based system for representing syntactic structure. Examples include the French MCVF (*Modéliser le changement: Les voies du français*) corpus (Martineau 2010), the

Portuguese *Tycho Brahe corpus* (Galves, de Andrade and Faria 2017), the Icelandic IcePaHC corpus (Wallenberg et al. 2011), the Irish POMIC corpus (Lash 2014), and the ongoing Corpus of Historical Low German (Koleva et al. 2017, Farasyn et al. 2018).

A second tradition has developed a dependency-based system for representing syntactic structure in early Indo-European languages. Corpora in this tradition that have adopted a common annotation system include the PROIEL Treebank (Haug and Jøhndal 2008), which includes texts in Ancient Greek, Latin, Gothic, Classical Armenian and Old Church Slavonic; the TOROT treebank (Eckhoff and Berdicevskis 2015) for Old Church Slavonic, Old East Slavic and Middle Russian; and the ISWOC treebank (Bech and Eide 2014) for Old English and early Romance languages.

This paper outlines ongoing work in the first of these traditions to create a constituency-based parsed historical corpus for Welsh, namely the Parsed Historical Corpus of the Welsh Language (PARSHCWL). The original Historical Corpus of the Welsh Language (HCWL) project (Willis and Mittendorf 2004a, b) compiled a collection of some 30 texts from the period 1500–1850, amounting to around 420,000 words, encoded in an XML format in conformity with the guidelines of the Text Encoding Initiative (TEI). PARSHCWL expands this collection back in time to include the whole of the Middle Welsh period (from 1250), and, above all, enhances it so that every word is identified for its part of speech and every sentence is assigned a constituent structure. This is a complex procedure that requires a robust workflow so that a team of researchers with a range of different linguistic and computational skills can work together. This paper outlines the issues involved in developing conventions for the corpus and the workflow procedure itself, beginning with an overview of all aspects of corpus preparation (section 2), before focusing in detail on the two first major stages, namely preparation of texts (section 3) and the part-of-speech tagging procedure (section 4). We conclude (section 5) with a discussion in somewhat more detail of some specific issues that arise in three case studies: word division, tagging of verbal inflection, and a tagging issue in relative clauses.

2 Developing an annotation procedure

2.1 Philosophy and goals

In line with this long-standing tradition of the development of parsed historical corpora, the goal of PARSHCWL is to facilitate automated searches of historical texts. Texts are taken from a variety of sources, some from the original HCWL corpus, some from other text-encoding projects or from editions. At each step of the preparation and annotation procedure, decisions have to be made with the ultimate goal in mind, namely to prepare the texts in ways that facilitate later annotation and to develop an annotation standard that will provide maximal assistance to the end user with acceptable outlay of effort on the part of the project team. The aim is thus not to provide a ‘correct’ linguistic analysis of each word and sentence within the corpus, but rather to provide as much morpho-syntactic information as possible, in order to enable linguists to then conduct their own searches and analyses based on their own hypotheses and research questions. Although syntactic structure is presented in the form of trees and bracketed representations, inclusion of hierarchical structure is limited and, whenever possible, straightforward and relatively theory-neutral phrasal labels such as ‘noun phrase (NP)’ or ‘prepositional phrase (PP)’ are employed.

In line with this, all morpho-syntactic labels added to the text have clear meanings: they were chosen to avoid presenting the user with ambiguous or unclear cases. The morphological (part-of-speech, POS) and syntactic categories chosen for PARSHCWL build above all on the tag sets developed for the English and Icelandic historical treebanks. This method was adopted in order to make it easier for researchers trained in the use of these corpora – the English corpora are often used in introductory training, for instance – and to allow researchers working on languages other than Welsh to make full use of the Welsh historical treebanks. These tag sets necessarily have to be adapted to reflect Welsh grammar; for instance, PARSHCWL is unique in providing additional information in the form of hyphenated tags specifying person, gender and number features on verbs and prepositions, since they represent an important part of Middle and Early Modern Welsh grammar. All decisions about the pre- and post-processing of the texts, as well as the annotation procedure and morpho-syntactic labels themselves are described in a detailed annotation manual that will be made available online. Some specific case studies, describing the problems that arise in choosing an appropriate tag set and constituency and hierarchical structure, and how these problems have been dealt with, are discussed in further detail in section 5.

The original HCWL corpus (Willis and Mittendorf 2004a, b) was balanced in the sense that it included roughly equal amounts of narrative prose, political/didactic prose, historiography, religious literature/scripture, personal letters, drama and free-metre poetry/ballads. PARSHCWL, however, is not yet a balanced corpus. There are more significant difficulties in achieving this kind of balance for the medieval period than for the early modern period, and for maintaining any text types identified over time. A well-balanced corpus would need to include a wider range of text types. This first attempt at creating a parsed historical corpus of Welsh contains only prose, since decisions about sentence boundaries and syntactic structure in poetry are notoriously difficult. In future, however, with the support of researchers working on Middle Welsh poetry, it may be possible to incorporate poetry into PARSHCWL.

2.2 Annotation procedure

Corpus annotation is a difficult and time-consuming task that has to be carried out with great precision and consistency. In order to optimize the workflow, a robust annotation procedure needs to be developed. The four steps of this procedure for PARSHCWL are pre-processing, POS-tagging, parsing, and post-processing, as shown in Figure 1. Each of these steps is further broken down and organized into a series of sub-tasks. This arrangement is designed to maximize efficiency when working with a team of annotators. Since each of the sub-tasks requires a different set of skills, this allows as large a number of people as possible to work collaboratively on the corpus, each using their own skillset, so that new texts can be added as quickly as possible to the growing parsed corpus while maximizing consistency across the treatment of different texts. Each of these steps is discussed in more detail below.

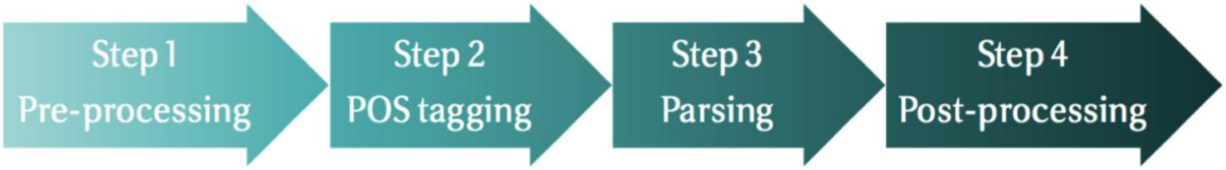

Figure 1. The four steps of the PARSHCWL annotation procedure.

Pre-processing is necessary to prepare all digitized texts for any type of automatic annotation, whether it is morphological (POS tagging) or syntactic (parsing). During the pre-processing step, texts are obtained that have been digitized from manuscripts, early printed texts or editions and/or made available by previous projects, such as the large collections of Middle Welsh manuscripts by Isaac and Rodway (2002), Luft, Thomas and Smith (2013) or Roberts, Rowles and Sims-Williams (2015), or Early Modern Welsh manuscripts and printed texts from the original HCWL corpus (Willis and Mittendorf 2004a, b). Text files are then edited to make them suitable for subsequent semi-automatic annotation, bearing in mind differences in transcription and encoding practice, for instance, treatment of spaces and line breaks. Texts are provided with a unique TextID, and each sentence is given a unique SentenceID so that search queries in the corpus can be reliably traced back to the source text. Finally, decisions are made to either split or join words, a process that is called 'tokenization' (see the case study in section 5.1 below). These changes involve a number of difficult decisions, and these are documented in an annotation log that other annotators can consult whenever they are confronted with a similar decision for which there is as yet no guideline. In the first stages of developing the PARSHCWL, difficult decisions were flagged so that they could be discussed at the next annotator meeting, at which a new guideline was written and added to the annotation manual. Pre-processed texts are all saved as text files in a UTF-8-encoded .txt format.

Once the texts are pre-processed, each word (token) can be provided with a morpho-syntactic label, in the form of a part-of-speech (POS) tag. POS tagging can be done automatically, but the results can never be 100% accurate. The POS-tagging step therefore also includes a round of manual correction in which a human annotator checks whether the algorithm has selected the correct tag for each token. Again, changes and difficult decisions are documented in an annotation log and discussed within the team of annotators whenever necessary. As is standard in the other corpora, POS-tagged texts are saved as text files with a .pos extension.

The corrected POS-tagged files are then automatically chunk-parsed, that is, provided with phrase-structure labels, such as NP (noun phrase) or PP (preposition phrase). This is done by a rule-based parser that combines words into groups based on their POS tags. An example of such a rule to combine determiners, nouns, adjectives, numerals and demonstratives into Noun Phrases (NPs) is given below:[2]

(1) NP: {<D>?<NUM|ONE>?<N>?<ADJP>?<DEM>?}

This rule says that if the chunk parser encounters a sequence of words tagged as D NUM N ADJP DEM (e.g. Middle Welsh y_D tri_NUM ty_N (ADJP da_ADJ) hyn_DEM 'these three good houses'), they can be combined into a unit labelled NP. The question marks indicate that the preceding element is optional, allowing subsets such as y_D tri_NUM hyn_DEM ‘these three’ or cath_N ‘a cat’ to satisfy the rule. The pipe indicates either/or, hence NUM|ONE means either a numeral (NUM) or the word *un* ‘one’ (ONE). With the NLTK RegEx chunk-parser, the above rule automatically converts a phrase like (2a) into a structured NP represented in the standard bracket notion used by parsed (.psd) files, shown in (2b) (on the referencing system, see below):

(2) a. *ẏ*/D *nos*/N *honno*/DEM
b. (NP (D *ẏ*)
(N *nos*)
(DEM *honno*))
‘that night’ (PKM 1.4)

Adapting the Natural Language Toolkit (NLTK) Python chunk-parsing script, we can also create further hierarchical structures to create fully parsed sentences. A simple example of a Noun Phrase within a Prepositional Phrase is given in (3), applying the rule in (4) (a rule similar to (1) above will have already created the NP from *y gydymdeithon*).

(3) a. *a*#/P *ẏ*/PRO-G *gẏdẏmdeithon*/NPL
b. (PP (P *a*)
(NP (PRO-G *y*)
(NPL *gydymdeithon*)))
‘with his companions’ (PKM 1.10)
(4) PP: {<P><NP>}

For further details and examples of chunkparsing, see Meelen (2016) and Meelen and Willis (forthcoming).

As with the POS-tagging step, the results of automatic parsing have to be manually corrected and changes and decisions are documented in an annotation log. The parsed versions of the text include both the POS tags as well as the phrasal tags and are saved as text files with .psd (‘parsed’) and .psdx (the equivalent in XML) extensions. These formats are chosen because they are compatible with various types of query software widely used to search this type of annotated corpus, for instance, CorpusSearch (Randall, Taylor and Kroch 2005) and CorpusStudio (Komen 2009b).

The final step is post-processing. During this stage, all files are checked to ensure that they are ready to use in the correct format with the names adhering to project file-naming conventions. Finally, metadata for each text are added to the files, providing further information about manuscripts or printed texts, modern editions, author and date of composition (where known). In principle, there is no limit to the amount of information that is added here, which affords us the flexibility to expand metadata entries at a later date. This metadata is available in the XML-versions of the parsed files, but can also be accessed separately through the project website. Further information about the tasks in each step of the described procedure is available there for future annotators.

Having outlined the overall architecture of the procedure, we will now turn to discuss two stages in more detail, namely pre-processing and POS-tagging.

3 Pre-processing

Before texts can be automatically provided with morpho-syntactic labels, they need to be prepared in various ways so that they match the required input format for the annotation software and can be incorporated into PARSHCWL. This section outlines the general procedure for preparing resources, before examining the two most important aspects of preparing the texts for further natural language processing, namely normalization and tokenization.

3.1 Preparation of texts

Texts for this first phase of PARSHCWL were selected based on two criteria: (i) texts were chosen for which digitized manuscript transcriptions were readily available; and (ii) texts were chosen so as to trial the text-preparation workflow with different text types. Since the addition of highly detailed morpho-syntactic annotation is a challenging task, we chose to focus initially on two major Middle Welsh narrative prose texts, namely *Culhwch ac Olwen* and *Pwyll Pendefig Dyfed*, in both their manuscript versions, the White Book of Rhydderch and the Red Book of Hergest, available at www.rhyddiaithganoloesol.caerdydd.ac.uk. The next part of the first phase of PARSHCWL will extend this to include selections from the Middle Welsh laws (from BL

Add. 22356, James 2013) and *Ystoria Adrian* from Llyfr yr Ancr (Jesus ms. 119, The Book of the Anchorite of Llanddewi Brefi) (Parina et al. 2018). For Early Modern Welsh and later periods, we used a selection of the texts digitized for the HCWL corpus (Willis and Mittendorf 2004a, b), namely extracts from *Cronicl Hywel ap Syr Mathew*, the 1588 Bible translation, Huw Lewys's *Perl mewn adfyd* (1595), Ellis Wynne's *Gweledigaetheu y Bardd Cwsc* (1703) and the letters of Goronwy Owen (1750s). Examples from these texts will be included here where appropriate. The procedure is designed in such a way that it can be extended to any type of text at any time, but these texts form an excellent starting point.

Selected texts were first of all added to a corpus database recording necessary metadata, such as the name and origin of the text and manuscript and important editions for further reference. In this overall corpus database, texts are also provided with a TextID, a file name that adheres to a convention for cataloguing all texts. For instance, the TextID PwyllWB refers to the text *Pwyll Pendefig Dyfed* (Pwyll) in its White Book of Rhydderch version (WB). The period to which the manuscript dates is identified by the suffix -m4, in this case identifying the manuscript text as coming from the fourth Middle Welsh period (-m4). In a procedure mirroring that adopted by the Middle English (PPCME) corpus, the Middle Welsh period is divided into five century-long periods, from the eleventh up to the end of the fifteenth century; -m4 thus refers to a text from a manuscript dating to the fourteenth century.

Once text files are consistently named and catalogued, they are stripped of any mark-up that will interfere with automated tagging and parsing (e.g. XML or HTML depending on the source), and saved as plain text files converted to the correct encoding type (UTF-8). Finally, texts are split into sentences, with each sentence receiving a unique sentenceID consisting of the textID plus a running sentence number starting from the beginning of the file. In the first instance, sentences are split up automatically based on whatever punctuation is available in the source: after each full stop, an utterance boundary tag <utt> is added automatically. However, punctuation is not always available in the manuscript; nor is it consistently used in Middle Welsh (see further Poppe 2018). A manual check is therefore necessary to ensure that sentence boundaries have been placed in useful and appropriate positions. As with any language, opinion can vary on the definition of a 'sentence'. For PARSHCWL, where a choice is available, the presumption is to choose to make sentences as small as possible, in order to minimize the task of subsequent correction of the morpho-syntactic annotation: shorter sentences produce less complicated representations, making it easier for annotators to make any necessary changes. Each sentence consists minimally of a main clause. A number of clause types are considered separate utterances in PARSHCWL. With co-ordinated clauses, as in (5), each conjunct is treated as a separate utterance.

(5) a. Mi a th rodaf di *ẏ# m lle# i ẏn annɓuẏn
I PRT 2SG give.1SG you to 1SG place 1SG in Annwfn
‘I’ll put you in my place in Annwfn…’ (PKM 3.8)
b. *ac a rodaf ẏ ɓreic deccaf
and PRT give.1SG the women fair.SUPERL
‘…and I’ll give you the most beautiful woman…’ (PKM 3.9)

Absolute clauses, non-finite adjunct clauses expressing circumstance introduced by *a*(*c*) ‘and’ (Richards 1938: 26–28, Williams 1980: 172–173, Rottet 2011: 240–241), as in (6), are split off in the same way, as are verbal noun clauses functioning as main-clauses, as with the second conjunct (*a dyuot…*) in (7).

(6) ac Arthur y ar Hengroen y uarch.
and Arthur from on Hengroen 3MSG horse
‘and/with Arthur [being] on Hengroen, his horse’ (CulhwchWB-m4-348)
(7) ...kyuodi a oruc, a dyuot y Lynn Cuch…
arise.INF PRT do.PST.3SG and come.INF to Glyn Cuch
‘He got up and came to Glyn Cuch…’ (PKM 6.16)

However, other subordinate clauses are always linked to a main clause, as with the temporal *pan*-clause in (8):

(8)	A	phan	doeth	yno,	yd	oed	Arawn	urenhin	Annwuyn
	and	when	come.PST.3SG	there	PRT	be.IMPF.3SG	Arawn	king	Annwfn
	yn	y	erbyn.						
	in	3MSG	against						

‘And when he came there, Arawn, king of Annwfn met him.’ (PKM 6.16-17)

3.2 Normalization

In order to make the manuscript versions of the text optimally readable, two normalization tasks are performed involving punctuation and orthography.

Punctuation is added and/or changed from the manuscript whenever it has been added by our source, but is not currently added to manuscript transcriptions.

Orthography is generally left as in the source, even the graph ỽ (Middle Welsh v, unicode U+1EFD. Barred double ỻ is also normalized to <ll>. Scribal errors are corrected in the text files to facilitate POS tagging and general legibility and interpretability. Where text is amended in any way (including scribal insertions), this is marked by adding an asterisk to the beginning of any affected words. This does not impact on POS-tagging (see below), but does alert the user to a possible issue with the text at this point, which, depending on the use being made of the corpus, may need to be investigated further by the researcher. The details of the emendation are not currently stored within the corpus: the corpus is not intended to be a fully self-standing resource, and should always be used in conjunction with existing resources, above all, electronic and traditional editions of the texts in question.

3.3 Tokenization

Tokenization is the task of splitting a string of characters up into meaningful words or ‘tokens’. This is necessary to ensure that part-of-speech tags are applied to meaningful units.

Modern and Early Modern Welsh manuscripts and printed texts generally use spaces to indicate word boundaries, but this is not always the case in Middle Welsh manuscripts. Even manuscripts that do employ spaces in this way do not always place word boundaries in places that correspond to meaningful words or grammatical units that linguists can work with.

For instance, *ẏnẏ uỽrỽ* ‘throwing him’ (PwyllWB-m4-9) is best tagged with *ẏn* as a progressive marker PROGR and *ẏ* as a prefixed genitive pronoun PRO-G. To do this, *ẏnẏ* needs to be split into two tokens. This avoids the need to create complex and difficult to handle composite tags such as PROGR+PRO-G, and allows users more easily to extract, say, all the genitive pronouns.

All texts are therefore automatically divided into tokens (tokenized) by identifying spaces (white space) in the first instance, but manual changes are made afterwards whenever the need to split, join or add tokens is identified. Further details of this procedure will be provided in section 5.1 below.

Once all the texts are pre-processed, they are saved as text files that are ready for subsequent part-of-speech tagging. During the manual correction of this step, annotators keep notes on changes and flag up difficult decisions in an annotation log. In this way, ambiguous or challenging cases can be discussed within the team of annotators to ensure a high level of consistency throughout the annotation process. Consistency is crucial as it allows users to be confident of how a given structure of interest will have been dealt with, allowing them to formulate predictable, and therefore accurate, searches. Decisions on specific cases are logged in a notes file, which will be available, along with the pre-processed texts files themselves, on the project web portal.

4 POS tagging

Once pre-processing is complete, the next stage is to label each token with an appropriate part-of-speech (POS) tag. This is illustrated for the first sentence of *Pwyll Pendefig Dyfed* in (9). Here, each of the first three words is identified as a proper noun (NPR), as is the final word *dẏuet*; *a* is a verbal particle (PCL); *oedd* is a third-person (3) singular (SG) imperfect (A) indicative (I) form of the verb *bod* ‘be’ (BE); *yn* is a predicate marker PRED; *arglỽẏd* and *cantref* are common singular nouns (N); *ar* is a preposition (P); and *seith* is a numeral (NUM). Punctuation is treated as a separate token and given the tag PUNC. This is to ensure that it is recognized and included in the structural tree representation later. Tags are separated from the word they modify using a forward slash (/). Finally, each sentence is identified by a full reference: the TextID is PwyllWB and this is the first sentence of the text (-1); it will be found at line 1 of Ifor Williams’ standard edition (PKM-l1) and on folio 1recto of the manuscript Peniarth 4 (i.e. the White Book of Rhydderch).

(9) Pwẏll/NPR pendeuic/NPR dẏuet/NPR a#/PCL oed/BEAI-3SG ẏn/PRED arglỽẏd/N ar/P seith/NUM cantref/N dẏuet/NPR ./PUNC PwyllWB-m4-1_PKM-l1_Pen4-f1r_CODE <utt>

There are two parts to defining a successful procedure for generating representations of the type in (9). First, the tag set itself must be defined. This is no trivial task. As noted above, the purpose of the tag set is not to provide a definitive analysis of Middle Welsh grammar, although it may point in a certain preliminary direction; rather, it functions as a search aid for the researcher. Furthermore, it will serve as the input to parser in the next stage of annotation. Different uses of a single word are therefore disambiguated where it is thought that this could be useful for a researcher or for the parser, but not where a user could easily search for a distinction by other means. On the one hand, *yn* can be a preposition 'in' (P), a predicate marker (PRED) or a marker of progressive aspect and related uses (PROGR) (see also Scherschel 2019). These different uses could not confidently be disaggregated by a user from a search simply for the word *yn*. On the other hand, the demonstrative *hwnn* is tagged as a demonstrative (DEM) in all uses; the tagging does not distinguish independent (pronominal) from adnominal use, nor does it identify when (*yr*) *hwnn* is being used as a relative pronoun; a list of these uses could easily be extracted using POS-tags and/or the structural parse, and the value of this distinction to the corpus therefore does not merit the annotator time that would be necessary to implement it. There is thus an inevitable trade-off between the granularity of the tag set and the time taken to tag a given text (and thus, ultimately, the size of the corpus that can be produced).

In addition, the tag set must be robust enough to provide a tag for all words in a text and to cope with linguistic variation between texts from different text types and from different time periods. This last factor is a particularly difficult one, since language change means that a tag set appropriate for Middle Welsh will not be appropriate in every detail for Modern Welsh.

The design of the tag set also takes into account the architecture of existing projects in the Penn tradition. The point here is that the tag set and the corpus should not present unnecessary difficulties for a researcher trained in the use of the other corpora. We cannot simply take over the tag set of one or other of these corpora, since the tag set must be appropriate to the grammatical distinctions made in Middle and Modern Welsh. Nevertheless, divergences between the corpora should be motivated and not simply arbitrary.

Secondly, the tag set must be applied to the texts being processed. For this to happen, the texts must be tagged using an automated tagging procedure, expanding on the blueprint developed and described by Meelen (forthcoming), and then corrected. In this case, we compared the results of the Memory-Based Tagger (MBT) (Daelemans et al. 1996, and for Middle Welsh specifically, Meelen forthcoming) with an off-the-shelf neural-network tagger (BiLSTM-CNN-CRF from Targer, Chernodub et al. 2019). A portion of the corpus is set aside as a training corpus, and tagged manually in accordance with the tagging schema. For the neural-network tagger, we furthermore created word vectors with FastText (Bojanowski et al. 2017) based on all digitized texts of the Mabinogion tales and romances, *Llyfr yr Ancr* and a number of medieval law manuscripts.

The tagger generalizes from this training corpus to new material, which is then itself corrected manually and used to re-train the tagger, whose accuracy improves with each iteration of the process. This is a form of supervised learning in which the tagger generalizes from the most similar cases that it has previously encountered. In this way, we can begin with a relatively small manually tagged corpus (in this instance, the Middle Welsh tale *Branwen* was tagged manually to set the procedure in motion). If the tagger meets a word that it has previously enountered, it assigns the previous tag (e.g. *arglꝩẏd* is tagged as N). If the word is known, but has received multiple tags in the past (as with *yn*, which may be PRED, PROGR, P or PSUB), the tagger uses context, the tags assigned to surrounding words, to choose the tag that is most likely to be correct for the given instance. For unknown, previously unencountered words, the tagger uses both the form of the word itself, specifically its final three characters, and the context to estimate the tag most likely to be correct. This means, for instance, that verbal endings are likely to allow a verb, along with its tense, mood, person and number, to be identified correctly, even if the particular verb in question has not previously been encountered. Proceeding in this way resulted in automated tagging with a global accuracy of 90.4% for the memory-based tagger and 87% for the neural-network tagger. The results of the latter can be significantly improved with better word embeddings based on a larger corpus. For both taggers, more manually corrected training data (termed ‘Gold Standard’ since it reflects the best possible answer) and reducing the number of different morpho-syntactic tags (now over 200) would improve the results too; but if we cannot or do not want to do either of these things then, in general, this is a reasonably good success rate for a historical language sample including extensive orthographic variation within it.

Once the tagger has tagged the text, the tagging is corrected manually by an annotator familiar with the tag set. Unlike much of the pre-processing, this requires both a secure knowledge of Middle Welsh and of the tag set and a high degree of concentration. Creating a highly structured workflow thus allows annotators' time to be focused most efficiently on the areas of the procedure where the skills are most in demand. In addition to correcting the tagging in accorance with the annotation manual, the annotator identifies any cases where the correct tag is not clearly identified in the manual. As in the pre-processing stage, such cases are logged and referred back to the whole team, where they are discussed and the tag set and annotation manual amended if necessary.

The output of POS tagging can, if desired, be converted automatically into other formats for users not coming to the corpus from a Penn-style tradition, for instance, the Leipzig glossing rules (Comrie, Haspelmath and Bickel 2015 [2008]), which may be more familiar to general linguists.

5 Pre-processing and POS tagging: Case studies

5.1 Separating words: splitting and joining tokens

As mentioned in section 2.2 above, tokenization entails the splitting of a string of characters into meaningful words or units. After automatic tokenization based on the white space as found in the manuscript versions of the text, manual correction is necessary to split, join or add words to make the annotated text more readable and ready for linguistic analysis. This section gives a detailed description of each of these tasks and why they are necessary to create a fully parsed historical corpus.

5.1.1 Splitting tokens

In many Middle Welsh manuscripts, clitics, prepositions or other short words (often consisting of 1 or 2 letters) are merged orthographically with a following or preceding word. To facilitate a consistent linguistic analysis, these words are split so that they each receive their own informative part-of-speech tag. In principle, all words (including particles and clitics) are thus split and tagged individually. Whenever a combination is split, the first part is consistently extended by a # sign to indicate that this particular word boundary deviates from the original found in the source. Common occasions of such splits include combinations of preverbal particles (PCL), conjunctions (CONJ) or complementizers (C) with infixed accusative object pronouns (PRO-A) or definite articles (D), as shown in Table 1.

Context	**Split tokens**	**POS tags**	**Translation**
Mi ae gwelaf	*a# e*	PCL PRO-A	‘I see her’
ar mab	*a# r*	CONJ D	‘and the boy’
peis	pei# s	C PRO-A	‘if … him/her/it/them’

Table 1. Common examples of word splitting in PARSHCWL.

In addition to preverbal particles that occur in verb-second contexts (in ‘abnormal sentences’), particles in sentence-initial position are split from their following verbal tokens in a similar way. This is shown in Table 2. Note that use of the tag PCL allows us to cope with historical change in the form of verbs and preverbal particles: Modern Welsh *r-* receives the tag PCL in the same way that Middle Welsh *yr* does. This does not commit us to this as the correct synchronic analysis, but means that both can be found by a researcher familiar with the corpus guidelines.

Context	**Split tokens**	**POS tags**	**Translation**
ydywch	yd# ywch	PCL BEPI-2PL	‘you are’
rydw	r# ydw	PCL BEPI-1SG	‘I am’

Table 2. Examples of splits involving preverbal particles.

Splitting and annotating tokens in this way is especially insightful in cases where multiple options and/or splits are possible. Examples are give in Table 3, where *os* may have any one of three POS analyses depending on the context.

Context	**Split tokens**	**POS tags**	**Translation**
os	[not split]	C	‘if’
os	o# s	C PRO-A	‘if … him/it’
os	o# s	C BEPI-3SG	‘if it is’

Table 3. Possible splits for Middle Welsh *os*.

Other common examples of merged tokens that are split include combinations with the preposition *yn* ‘in’ or the reflexive *ehun* ‘himself’ (or variants thereof), as shown in Table 4.

Context	**Split tokens**	**POS tags**	**Translation**
ygKaerfyrddin	yg# Kaerfyrddin	P NPR	‘in Carmarthen’
ehun	e# hun	PRO-G REFL	‘himself’

Table 4. Examples of splits with *yn* ‘in’ and *hun* ‘self’.

Sometimes, scribes merge more than two elements together, in which case multiple splits are inserted, as with *onys* in Table 5.

Context	**Split tokens**	**POS tags**	**Translation**
onys	o# ny# s	C PCL-NEG PRO-A	‘if not... him’

Table 5. Example of a multiple split with *onys*.

In certain cases, merger of words leads to apparent ‘deletion’ of letters shared by both elements. Such deleted letters are not reinstated, but are indicated with a + symbol on the first element. This is particularly a problem with the preposition *yn* ‘in’, which, for understandable phonological reasons, is normally written together with the following word in Middle Welsh, often in ways that represent the final consonant of the preposition and the initial consonant of the following word as a single unit (see Watkins 1968 for further details). Splits in combination with *yn* are made after the vowel to leave the initial consonant mutation of the following word in tact and to allow the + symbol to indicate the exact location of the missing final nasal of the preposition. This is mainly done to create consistency among the set of additional symbols, # and +: both are used only at the end of a word, and this convention facilitates future queries involving searches at the word level of the corpus. Examples are given in Table 6. Note that the merged and subsequently split negative *kanys* in Table 6 can now be distinguished from the non-negative complementizer *kanys* meaning ‘because’.

Context	Split tokens	POS tags	Translation
ymhrytain	y+ mhrytain	P NPR	‘in Britain’
kanys	ka+ ny# s	C PCL-NEG PRO-A	‘since not … it/him/her/them’

Table 6. Examples of use of ‘+’ to indicate split with loss of an element.

5.1.2 Joining tokens

In many of the texts under consideration, conventions for splitting words do not conform to present-day usage, and what we would today normally analyse as a single linguistic token, such as an inflected preposition (P-) or an adverb (ADV), can be split in two by a space. In other cases, even where they are conventionally spelled as multiple words today, fixed collocations or names whose internal structure is not obvious or compositional are combined in order to avoid the needed to create separate tags for each element within them. To facilitate future research on such tokens and to enable a common search procedure for all texts, such instances are combined into a single unit in PARSHCWL. Tokens that are combined in this way are marked with a ! symbol in between the parts to indicate the split as found in the original source. Sometimes, more than one separate word in the source is combined in this way. A range of examples is given in Table 7.

Context	Joint tokens	POS tags	Translation
o honi	o!honi	P-3SGF	‘from her’
am danaw	am!danaw	P-3SGM	‘about him’
nid amgen	nid!amgen	ADV	‘namely’
hyd yn nod	hyd!yn!nod	ADV	‘even’

Table 7. Examples of the use of the joining element ‘!’ in PARSHCWL.

Combining tokens in this way has several advantages. First of all, it simplifies the subsequent POS tagging and parsing steps of the annotation procedure, since there is no need to invent additional tags to represent the original split. Adding new morphological tags to the tag set is to be avoided unless there is a meaningful morphological distinction, since it severely impacts on the results of automatic POS tagging. The POS tagging algorithm would take longer to provide each token with a tag and since a larger tag set means there are more tags to choose from, adding tags would unnecessarily increase the chances of choosing the wrong one.

Another advantage of combining tokens is seen in the parsing step. A smaller tag set also means there are fewer possible combinations for the regular-expression chunk-parser that combines words into phrases on the basis of their morphological tag. The regular-expression grammar can be significantly simplified if there are fewer tags and it will therefore be less prone to error as well. For both these steps, combining words like these consistently therefore means less work during the stage of manual corrections. Finally, combining words consistently into meaningful tokens facilitates future research on these units. Use of the exclamation mark between the combined parts means that no information from the source is lost; nevertheless, the token can also be found as the expected unit on the basis of its POS tag. Inflected prepositions, for example, are always tagged ‘P-’ regardless of how they are found in the manuscript. A simple query for all tokens tagged ‘P-’ will yield all correct results in their respective contexts.

5.1.3 Adding tokens

A final case where the original text in the source is edited concerns additions of new tokens for words are present semantically but ‘disappear’ for phonological reasons. This often happens in Middle Welsh manuscripts with infixed pronouns combined with prepositions as they could be merged. Whenever tokens seem to have been merged in this way, the convention adopted for lost characters above is adopted, and a + sign is added in place of the ‘missing’ element and a POS-tag is assigned to the second element. This is illustrated in Table 8, where the tag PRO-G is assigned to the inserted element +, which is itself treated as a separate word (token), to indicate the presence of a prefixed genitive pronoun.

Context	Added tokens	POS tags	Translation
y dy	y + dy	P PRO-G N	‘to his house’

Table 8. Example of the treatment of inserted elements in PARSHCWL.

In many cases, the missing element is evident from the initial-consonant mutation on the subsequent word, as with the nasal mutation on *nghoelio*, indicating the presence of a prefixed first-person singular genitive pronoun in Table 9.

Context	Added tokens	POS tags	Translation
nghoelio	+ nghoelio	PRO-G VN	‘believe me’

Table 9. Treatment of missing element that triggers mutation.

Note that empty categories that are part of some syntactic theories (e.g. Minimalism), such as null subjects, traces or PRO are not indicated in the POS-tagged files, but some are introduced into the parsed files (.psd and .psdx). The + sign signals only actual words that usually have phonological content, but are not written in the source. Finally, after all tokenization tasks are complete, as with all steps of the procedure involving manual correction, a decision log is kept to facilitate discussion of difficult cases within the project team.

5.2 Verbal inflection

The tag set for verbal inflection provides a relative straightforward illustration of how the tagging works, the purpose of tagging, and some of the difficulties and compromises involved in creating a useable tag set for the history of Welsh.

First of all, it should be noted that the verbal tags in PARSHCWL are relatively rich. Unlike the other Penn historical corpora, PARSHCWL includes a tag indicating person–number features for all finite verbs and inflected prepositions. Given verbal irregularities and syncretisms within paradigms, these features cannot reliably be derived from the form of the verb, particularly for texts before the advent of standardized spelling. The POS-tagger is able to extract these relatively accurately, and it was felt that such tagging could be of use in research.

Penn corpora typically distinguish certain very common verbs used as auxiliaries. It was clearly useful for Welsh for forms of *bod* ‘be’ and *gwneuthur* ‘do’ to be distinguished, as these can be used as auxiliaries, and *bod* in particular manifests many forms and variants. These are marked using the BE tag and DO tag respectively. The tags HV ‘have’ and MD ‘modal’ are not used, Welsh having no direct equivalent to English *have*, and lacking a syntactically significant category of modal (see Borsley, Tallerman and Willis 2007: 44–47); however, GT is used for forms of *cael* ‘get’, which is used as a passive auxiliary, and is sufficiently common that it was felt it could be useful to allow all uses of it to be easily extracted. Some corpora add tags for verbs that are in some way special in the language in question: the French MCVF corpus has a tag unique to *aller* ‘go’, for instance. For Welsh, an addition tag BL (for ‘belong’), whose morphology and syntax is highly idiosyncratic, and which is not etymologically simply a verb, is added to deal with forms of *piau* ‘belong’.

A more difficult question concerns the number of verbal paradigms to be distinguished in the tagging. The system adopted essentially follows that of Evans (1964), with some adaptations to deal with difficult cases.

The structure of the verbal tag is designed to allow extraction of some natural categories using wildcard searches. All lexcial verbal tags begin with V (so V* will extract all lexical verbs). Nonfinite lexical verbs (verbal nouns) are distinguished as VN, while finite verbs are VB (so all lexical finite verbs can be extracted as VB*). Within the category of finite verbs, tense and mood are distinguished. The next two characters indicate the tense–mood–aspect value of the verb: PI for present indicative; PS for present subjunctive; D for past (preterite) (from the English -*ed* suffix); P for perfect (in those few verbs that distinguish perfect from preterite); G for pluperfect (from Welsh *gorberffaith*); AI for imperfect indicative (from Welsh *amherffaith*); AS for imperfect subjunctive; F for future (in those verbs where the future and present are morphologically distinct); and C for conditional (‘consuetudinal past’/habitual past) (again, only for those verbs which distinguish this from the imperfect subjunctive). Finally, a person–number suffix is added at the end of the verbal tag -1SG, -2SG etc., with -4 used for the impersonal forms. A full tag can thus take a form such as VBPS-1PL to be read as ‘lexical verb in the present subjunctive, first person plural’.

The maximal system is illustrated for the verb *bot* ‘be’ in Table 10. There is no contrast for *bot* between preterite and perfect, so the form *bu* is tagged as preterite (BED-3SG). For a verb like *mynet* ‘go’, where there is a contrast, a distinction in tagging is made between preterite *aeth* (VBD-3SG) and perfect *ethyw* (VBP-3SG). This system is designed to correspond to that in Evans (1964), allowing users to consult this reference work for guidance. Importantly, it also allows all verb forms to receive a unique tag.

Context	POS tag	Description
bod	BEN	verbnoun
bit	BEI-3SG	imperative
mae	BEPI-3SG	present indicative
yssyd	BEREL-3SG	relative present indicative
panyw, *y mae*	BEFOC-3SG	focus
bo	BEPS-3SG	present subjunctive
bu	BED-3SG	preterite
ethyw	VNP-3SG	perfect
buassei	BEG-3SG	pluperfect
oed	BEAI-3SG	imperfect indicative
bei	BEAS-3SG	imperfect subjunctive
canei	VBA-3SG	ambiguous imperfect
byd	BEF-3SG	future
bydei	BEC-3SG	habitual past / conditional

Table 10. Maximal tag set for Middle Welsh verbs.

One difficulty concerns the imperfect, where many forms are not unambiguously either indicative or subjunctive, particularly in later texts. While it would be possible to ask annotators to judge whether indicative or subjunctive was intended on any given occasion, this would clearly amount to the annotator involving themselves in analysis that should be carried out by a researcher using the corpus. For this reason, where a form cannot be uniquely identified as imperfect indicative or subjunctive, the tag can be left unspecified, as VBA.

Tagging is by form rather than function: a verb is tagged present (PI or PS) if it is formed according to the present tense paradigm, even if it has future meaning in context. While this means that the tags may become less semantically appropriate for later texts, individual paradigms can be reliably located and the interpretation left to the researcher. The principle that the tagging is an aid to searching and not an analysis is paramount here. Thus the *buasai* paradigm of the verb ‘be’, which merges in function with the *byddai* paradigm, is always glossed as pluperfect (BEG) in contrast to the conditional (BEC) of the latter, even if it is suspected that the language underlying the text makes no distinction.

5.3 Relative clause and the FREL tag

A more complex case illustrates the sorts of decisions that have to be made in difficult cases and the criteria used to resolve them. Middle Welsh relative clauses begin with one of the particles *a* or *y*(*r*). These are tagged as PCL in relative clauses, as elsewhere – the parsed version of the corpus will identify the relative clause itself, marking the structure as CP-REL (a complementizer phrase, i.e. a clause, of the type REL) and identifying the position relativized on (subject relative, object relative etc.). These particles also occur in free relatives, and this use also presents rather little difficulty, as the particles can still be tagged as PCL, and the free relative itself will subsequently be identified in the structural description as CP-FRL:

(10) a ’r ker6yneu oll, ac a vo danunt o ’r dillat g6ely.
and the barrels all and prt be.PRS.SBJV.3SG under.3PL ofthe clothes bed
‘and all the barrels, and whatever of the bedclothes may be under them.’ (Machlud Cyfraith Hywel, unit 522)

(11) a#/CONJ r/D kerꝩyneu/NPL oll/Q ,/PUNC ac/CONJ a/PCL vo/BEPS-3SG danunt/P-3PL o#/P r/D dillat/N gꝩely/N ;/PUNC Laws-m5-844_James2013-l844_CODE <utt>

The same free relatives, however, participate in a number of other constructions where the tagging is less obvious. They are often preceded by *or* with a partitive reading. Hence, in example (12), quoted here from Ifor Williams’ edition, the free relative has the sense ‘of everything that he had seen of the world’s hunting dogs’ or ‘of whatever he had seen of the world’s hunting dogs’.

(12) Ac o’r a welsei ef o helgwn y byt, ny welsei
and OR PRT see.PLUPERF.3SG he ofhunting.dogs the world neg see.PLUPERF.3SG
cwn un lliw ac wynt.
dogs same colour as 3PL
‘And of all of the hunting dogs he had ever seen, he had not seen any dogs the same colour as them.’ (PKM 1.20–21)

This raises the question of how to tag *or* here. Note, first of all, that, while the manuscript has *or*, Williams' edition, like many other editions of Middle Welsh texts, regularizes this to *o'r*, indicating an analysis with *o* as a preposition, plus a second element *'r*. D. Simon Evans (1964: 70) refers to this as a demonstrative, which would lead us to tag as DEM, although, etymologically, as Evans notes, it is probabaly a form of the definite article, which suggests a tag of D (determiner). Indeed, Lewis and Pedersen (1937: 220) simply treat it as the definite article. We also find related forms, *oc*, *ar* and *ag* in the same syntactic environment, particularly in later texts:

(13) ac am bob lle oc y dylyei hayarn uot arnunt y bydei
and at every place OC PRT should.IMPF.3SG iron be.INF on.3PL PRT be.COND.3SG
eur o gỽbyl.
gold of all
'and at every place where there should be iron on them, there would be gold completely.' (Jesus 111 Red Book of Hergest [*Math*], p. 186r, col. 2, ll. 23–24)

(14) canys pawb ar a gymmerant gleddyf, a ddifethir â chleddyf.
for everyone AR PRT take.PRS.3PL sword prt destroy.PRS.IMPERS with sword
'for everyone that takes a sword will be destroyed by a sword.' (1588 Bible, Matthew 26:52)

(15) Y mae *Cywydd y Farn* a 'r Nodau gorau ag a fedrwn wneuthur
PRT be.PRS.3SG Cywydd y Farn and the notes best AG PRT can.IMPF.1SG do.INF
wedi mynd i Allt Fadawg
PERF go.INF to Allt Fadog
'Cywydd y Farn and the best notes that I could make have gone to Allt Fadog…' (Letters of Goronwy Owen, NLW ms. 17B, p. 238)

How are we to tag *oc* and *ar* in these examples? The overriding considerations here are the need to ensure ease of searching, the need to be consistent, and the need to be able to deal with as wide a range of texts as possible, including both Middle and Early Modern Welsh texts. Our tagging scheme should be able to deal with all of these forms, and enable a user to extract all of them with ease if this is their focus of interest. Consequently, we should not be using radically different tags for quite similar constructions in the three cases unless we think there is some distinction that users would find useful. We should also try to remain as neutral as possible in terms of analysis. In this case, it was felt that to use either the tag DEM or D for *-r* in (12) would be problematic for two reasons: first, it prejudges the analysis, since it is far from self-evident that *-r* is a demonstrative or a determiner; secondly, and more importantly, it would be difficult to apply this to all the cases under consideration: would *ar* also be tagged in this way? What about *oc* in (13)? Furthermore, searching the corpus would be made easier if all of these cases were identified by a common tag that could be searched for. This is especially the case because the sequences of characters involved are common in Middle Welsh and frequently have some other interpretation: *o'r* can be preposition + article, *ar* can be a preposition, *oc* can be a form of the preposition *o* before certain pronouns etc. Finally, we cannot rely on the parse to identify the relevant cases (via the CP-FRL node) since, while these are free relatives in origin, the elements are also found in other contexts, particularly in later periods.

For this reason, a distinct tag, FREL (free relative marker) was created for these cases, and the two types, with and without *o*, are distinguished as P FREL and FREL respectively. Where *oc* appears instead of *or*, *-c* is treated as a variant of *-r*, hence tagged as FREL. The decisions made are summarized in Table 11. FREL is not intended as an analysis, merely as a means of extracting cases that could then be analysed separately.

Context	**Split tokens**	**POS tags**
or	*o# r*	P FREL
oc	*o# c*	P FREL
ar	*ar*	FREL
ag	*ag*	FREL

Table 11. Treatment of free-relative markers in the tagging system.

The various examples (12)–(15) above would thus be tagged as follows (with the relevant items in bold):

(16) ac/CONJ **o#/P r/FREL** a#/PCL ỽelsei/VBAI-3SG ef/PRO o#/P helgỽn/NPL ẏ#/D bẏt/N ./PUNC nẏ/PCL-NEG ỽelsei/VBAI-3SG cỽn/NPL un/ONE lliỽ/N ac#/P ỽẏnt/PRO ./PUNC

(17) ac/CONJ am#/P bob#/Q lle/N **o#/P c/FREL** y/PCL dylyei/VBAI-3SG hayarn/N uot/VN arnunt/P-3PL y#/PCL bydei/BEC-3SG eur/N o/P gỽbyl/Q ./PUNC

(18) canys/C pawb/QPRO ar/FREL a/PCL gymmerant/VBPI-3PL gleddyf/N ,/PUNC a/PCL ddifethir/VBPI-4 â/P chleddyf/N ./PUNC

(19) Y/PCL mae/BEPI-3SG Cywydd!y!Farn/NPR a/CONJ ’r/D Nodau/NPL gorau/ADJS **ag/FREL** a/PCL fedrwn/VBA-1SG wneuthur/VN wedi/PERF mynd/VN i/P Allt!Fadawg/NPR

6 Conclusion

Historical treebanks have been created for a number of languages; however, while many texts have been digitized and made available for historical Welsh, there is as yet no parsed corpus. This paper has set out a procedure for the creation of the Parsed Historical Corpus of the Welsh Language. Each step has been outlined from pre- to post-processing and the various tasks and necessary tools in the pre-processing and POS-tagging steps have been described in some detail, focussing on three case studies: tokenization, annotation of verbal morphology and a newly developed tag specific for Welsh, FREL for the free relative marker.

We have seen how, during the pre-processing stage, different source texts for Middle, Early Modern and Modern Welsh prose are selected and prepared. The tokenization procedure ensures that texts are prepared so as to produce the optimal results in subsequent POS tagging and parsing steps, while nevertheless ensuring that the orthographical details of the original source are left in tact as far as possible. A POS-tagger subsequently trained on the basis of the manually corrected training-set Gold Standard gave over 90% global accuracy. This is a robust result for a small diachronic corpus with highly inconsistent orthography that was minimally pre-processed in order to preserve as much historical detail of the language as possible. Once this tagging has been checked by trained annotators, sentence IDs, consistent file names and relevant useful metadata are added.

During the part-of-speech tagging and parsing stages, all pre-processed texts are automatically provided with detailed morpho-syntactic tags. The results are then manually corrected. Once the texts are corrected, a rule-based regular-expression grammar created for the Middle Welsh Mabinogion corpus by Meelen (2016) can be applied to generate theory-neutral shallow hierarchical phrase structure. This was illustrated briefly above; the details of parsing and post-processing, steps 3 and 4 of the procedure in Figure 1, will be discussed in a follow-up article (Meelen and Willis forthcoming).

This type of searchable annotated corpus is not only useful for researchers interested in the history of the Welsh language; such corpora have become an essential tool for comparative linguists working on morphology, syntax, information structure and patterns of language change. The development of this parsed corpus is furthermore needed because Welsh is interesting from a typological point of view. It exhibits a number of morpho-syntactic features that are not shared with neighbouring Indo-European languages and a collection of historical text in particular can shed new light on this. A parsed historical corpus of Welsh will give a wide range of linguists the opportunity to conduct comparative research in each of those areas. The procedure developed here, along with detailed guidelines available in the annotation manuals on the PARSHCWL website, forms an solid blueprint that can be used for future annotation of further texts that can be added to the corpus.

References

Texts and corpora

Bech, K. and Eide, K. 2014. *The ISWOC corpus*. Oslo: Department of Literature, Area Studies and European Languages, University of Oslo. http://iswoc.github.io, accessed 10 July 2019.

Eckhoff, H. M. and Berdicevskis, A. 2015. Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank. *Scripta & e-Scripta* 14–15, 9–25.

Galves, C., Leal de Andrade, A. and Faria, P. 2017. *Tycho Brahe Parsed Corpus of Historical Portuguese*, www.tycho.iel.unicamp.br/~tycho/corpus/texts/psd.zip, accessed 4 June 2018.

Helsinki Corpus of English Texts. 1991. Compiled by M. Rissanen (Project leader), M. Kytö (Project secretary); L. Kahlas-Tarkka, M. Kilpiö (Old English); S. Nevanlinna, I. Taavitsainen (Middle English); T. Nevalainen, H. Raumolin-Brunberg (Early Modern English). Helsinki: Department of Modern Languages, University of Helsinki.

Isaac, G. and Rodway, S. eds. 2002. Rhyddiaith Gymraeg o Lawysgrifau'r 13eg Ganrif, Fersiwn 2.0: A searchable version of all the Welsh prose in 13th-century manuscripts. Reformatted by Silva Nurmio, Kit Kapphahn and Patrick Sims-Williams (2010); emended by Simon Rodway and Patrick Sims-Williams (2013). http://hdl.handle.net/2160/11163, accessed 4 June 2018

James, C. 2013. *Machlud Cyfraith Hywel: Golygiad o BL Add. 22356*. Cambridge: Seminar Cyfraith Hywel.

Kroch, A. and Taylor, A. 2000. *The Penn-Helsinki Parsed Corpus of Middle English (PPCME2)*. Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 4. www.ling.upenn.edu/ppche-release-2016/PPCME2-RELEASE-4, accessed 4 June 2018.

Lash, E. 2014. The Parsed Old and Middle Irish Corpus (POMIC). Version 0.1. https://www.dias.ie/index.php?option=com_content&view=article&id=6586&Itemid=224&lang=en, accessed 18 July 2019.

Luft, D., Thomas, P. W. and Smith, D. M. eds. 2013. Rhyddiaith Gymraeg 1300–1425. http://www.rhyddiaithganoloesol.caerdydd.ac.uk, accessed 4 June 2018.

Martineau, F. 2010. MCVF Corpus of Historical French. University of Ottawa. www.arts.uottawa.ca/voies, accessed 10 July 2019.

Parina, E., Sackmann, R. and Meelen, M. 2018. PARSHCWL: The annotated texts of *Llyfr yr Ancr*. Cambridge Apollo Repository. DOI: doi.org/10.17863/CAM.3

PKM - *Pedeir Keinc y Mabinogi*, ed. I. Williams. Cardiff: Gwasg Prifysgol Cymru, 1930.

Roberts, R. G., Rowles, S. and Sims-Williams, P. 2015. *Rhyddiaith y 15eg Ganrif: Fersiwn 1.0*. Aberystwyth: Department of Welsh, Aberystwyth University. http://hdl.handle.net/2160/26752, accessed 17 July 2019.

Wallenberg, J. C., Ingason, A. K., Einar Freyr Sigurðsson and Eiríkur Rögnvaldsson. 2011. Icelandic Parsed Historical Corpus (IcePaHC). Version 0.9. www.linguist.is/icelandic_treebank, accessed 4 June 2018.

Willis, D. and Mittendorf, I. 2004b. *A Historical Corpus of the Welsh Language 1500–1850.* http://people.ds.cam.ac.uk/dwew2/hcwl/menu.htm>, accessed 7 July 2019.

Secondary literature

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Borsley, R. D., Tallerman, M. and Willis, D. 2007. *The syntax of Welsh*. Cambridge: Cambridge University Press.

Comrie, B., Haspelmath, M. and Bickel, B. 2015 [2008]. Leipzig glossing rules. Leipzig: Max Planck Institute for Evolutionary Anthropology and University of Leipzig. https://www.eva.mpg.de/lingua/resources/glossing-rules.php, accessed 10 July 2019

Daelemans, W., Zavrel, J., Berck, P. and Gillis, S. 1996. MBT: A memory-based part of speech tagger-generator. In Proceedings of the Fourth Workshop on Very Large Corpora (WVLC-4), Copenhagen, Denmark. arXiv preprint cmp-lg/9607012.

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 of Computational Linguistics (ACL 2019)*. Florence: Association of Computational Linguistics.

Evans, D. S. 1964. *A grammar of Middle Welsh*. Dublin: Dublin Institute for Advanced Studies.

Farasyn, M., Walkden, G., Watts, S. and Breitbarth, A. 2018. The interplay between genre and syntax in a historical Low German corpus. In R. J. Whitt (ed.), *Diachronic corpora, genre, and language change*, 281–300. Amsterdam: John Benjamins.

Haug, D. T. T. and Jøhndal, M. L. 2008. Creating a parallel treebank of the Old Indo-European Bible translations. In C. Sporleder and K. Ribarov (eds.). Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008). Marrakech: Association for Computationa Linguistics, 27–34. http://www.lrec-conf.org/proceedings/lrec2008/workshops/W22_Proceedings.pdf, accessed 10 July 2019. http://www.lrec-conf.org/proceedings/lrec2008/workshops/W22_Proceedings.pdf

Koleva, M, Farasyn, M., Desmet, B., Breitbarth, A. and Hoste, V. 2017. An automatic part-of-speech tagger for Middle Low German. *International Journal of Corpus Linguistics* 22, 108–141.

Komen, E. 2009b. *Corpus Studio manual*. Nijmegen: Radboud University Nijmegen.

Kytö, M. 1996. Manual to the diachronic part of the Helsinki Corpus of English Texts: Coding conventions and lists of source texts, 3rd edn. Helsinki: Department of English, University of Helsinki. http://clu.uni.no/icame/manuals/HC/INDEX.HTM, accessed 10 July 2019.

Lewis, H. and Pedersen, H. 1937. *A concise comparative Celtic grammar*. Göttingen: Vandenhoeck & Ruprecht.

Meelen, M. 2016. *Why Jesus and Job spoke bad Welsh: the origin and distribution of V2 orders in Middle Welsh*. Utrecht: LOT Publications.

Meelen, M. forthcoming. Middle Welsh: POS tagging and chunk-parsing a partial corpus of native prose. In E. Lash, F. Qiu and D. Stifter (eds.), *Corpus-based approaches to morphoysyntactic variation and change in medieval Celtic languages*. Berlin: De Gruyter.

Meelen, M. and Willis, D. forthcoming. Towards a Welsh historical treebank, Part II: Parsing and post-processing.

Poppe, E. 2018. Patterns of Welsh punctuation from manuscript to print, 1346–1620: A pilot-study of the Annunciation narrative. *Studia Celtica* 52, 123–136.

Randall, B., Taylor, A. and Kroch, A. 2005. *CorpusSearch 2*. Philadelphia: University of Pennsylvania.

Richards, M. 1938. *Cystrawen y frawddeg Gymraeg*. Cardiff: Gwasg Prifysgol Cymru.

Rottet, K. J. 2011. Conjunctive pronouns in Modern Welsh: A preliminary corpus-based study. In K. Jaskuła (ed.), *Formal and historical approaches to Celtic languages*, 233–253. Lublin: Wydawnictwo Katolicki Uniwersytet Lubelski.

Scherschel, Ricarda. Middle Welsh *yn* in verbal noun phrases. *Studia Celto-Slavica* 10, 111–136

Watkins, T. Arwyn. 1968. Dulliau orgraffyddol Cymraeg Cynol o ddynodi'r treiglad trwynol. *Bulletin of the Board of Celtic Studies* 23, 7–13.

Williams, S. J. 1980. *A Welsh grammar*. Cardiff: University of Wales Press.

Willis, D. and Mittendorf, I.. 2004a. Ein historisches Korpus der kymrischen Sprache. In E. Poppe (ed.), *Keltologie heute: Themen und Fragestellungen*, 135–142. Münster: Nodus.

---

[1] The authors would like to thank Elena Parina and Simon Rodway for their assistance with this article. The research was funded by a British Academy–Leverhulme Trust Small Research Grant 'Developing a Welsh Historical Treebank' awarded to both authors, and

---

by a British Academy Postdoctoral Fellowship awarded to Meelen. The support of our funders is gratefully acknowledged.

[2] This rule has been adjusted somewhat for the purposes of exposition.