Marieke Meelen

2 Annotating Middle Welsh: POS tagging and chunk-parsing a corpus of native prose

1 Introduction

For a study on syntactic changes in Middle Welsh (see Meelen 2016), it was desirable to compile a searchable corpus of Middle Welsh, at least a partial one including the most important narrative literature from the medieval period. This chapter presents this first annotated corpus of historical Welsh. For the present study, a selection of Middle Welsh texts was used, based on their popularity among Welsh philological and linguistic researchers (the tales of the *Mabinogion*) and a less-researched excerpt of a law text (the *Laws of Women*) to compare the results in a different genre. This selection forms a first step towards the creation of a much larger, and well-balanced, Parsed Historical Corpus of the Welsh Language (PARSHCWL, see Meelen and Willis forthc.).

The *White Book of Rhydderch* and the *Red Book of Hergest* manuscripts, both dating from the 14th century, contain the most famous collection of Middle Welsh native¹ literature: the *Mabinogion*. These tales (of unknown authorship) derive in part from an oral literary tradition. They are thought to have their origin in the early medieval period but were only put down in writing several centuries later (see, among others, Davies 1998). For the present chapter, all extant tales of the *Mabinogion* (11 in total) were annotated, representing the narrative prose of the Middle Welsh period of the language, c. 1150–1500 AD. High-definition photographs of both of these manuscripts are available online via the websites of the National Library of Wales (www.llgc.org.uk – *White Book manuscript* Peniarth 4–5) and Jesus College, University of Oxford (<https://digital.bodleian.ox.ac.uk/> – *Red Book manuscript* Jesus College 111). The earliest text for the present corpus is an excerpt from a law text mainly concerned with the rights of women. This excerpt of the Early Welsh laws is taken from the BL manuscript Add. 22356 (S), one of the most important manuscripts in the tradition of the Welsh *Law of Hywel*. This *Law of Hywel* was the system of law in use in medieval Wales, based on an

¹ The term ‘native’ refers to the texts that are assumed to be originally composed in Middle Welsh, as opposed to ‘translated’ texts that are translations from other languages, such as chronicles translated from Latin.

older custom system. It is named after Hywel Dda ('Hywel the Good'), a 10th-century Welsh king. The British Library Law manuscript Add. 22356 (S), however, is from the mid-15th century. As with all Middle Welsh law texts, although the manuscripts are mostly dated from the late medieval period, (parts of) the texts go back many centuries (Davies 1966). The latest edition is accessible online via www.cyfraith-hywel.org.uk.

Although many of these manuscripts are available as digital photographs, not all of them have been converted to (online) editions with search options that facilitate philological and linguistic research. A large collection of Middle Welsh texts, including the texts of the *Mabinogion* included in the present study, is now available online through the Welsh Prose project (www.rhyddiaithganoloesol.caerdydd.ac.uk), but the search function is limited, and the texts are not annotated in any way. Searchable corpora are indispensable tools in historical linguistic research, both from a quantitative and qualitative perspective. Qualitative corpus research, for example, comprises investigations of the distribution of the different forms and constructions that are attested in a wide variety of texts. When analysing large amounts of texts or corpora in this way, linguists need to be extremely consistent in their approach to get meaningful and testable results. Categorising and labelling forms or structures in large amounts of data can be prone to error, because even the most careful researcher can change their ideas about the features and characteristics they use to disambiguate categories as they are confronted with more and more material. As the number of texts in need of investigation grows, it is no longer feasible to simply read and make notes. A further disadvantage of manual notes is that the results are much harder to verify and replicate, which is problematic in quantitative studies in particular. But qualitative studies can also benefit from searchable (annotated) corpora, as with improved computational methods specific or rare forms under investigation (for the purposes of philological research or comparative reconstruction, for example) are much easier to find. Therefore, it is useful to employ methods from the field of Natural Language Processing (NLP) and the tools created by computational linguists to build annotated searchable corpora. Because of their computational nature, these tools are designed to consistently deal with large amounts of data in a very short period of time. The results are consistent, following strict rules that are well-described in the annotation manuals (see Meelen and Willis forthc.), and can then be made readily available for any (Welsh) linguist.

Having said this, however, as an inflected language without a standardised orthography, Middle Welsh poses some specific challenges for ready-built NLP tools like part-of-speech (POS) tagging and parsing algorithms that automatically add morphosyntactic information (see Sections 3 and 4). Initial consonant mutation, for example, can yield a word like *pawb* 'everyone' in three different

ways: *pawb* ‘everyone’ (no mutation), *a phawb* ‘with everyone’ (aspirate mutation) and *i bawb* ‘to everyone’ (soft mutation). Furthermore, with five tenses (Present, Past, Pluperfect, Present Subjunctive and Conditional/Imperfect), seven different person-number-gender suffixes, various sets of pronouns and clitics and a wide range of functional particles, Middle Welsh with its rich morphology and extensive orthographical variation is far more difficult to automatically annotate than a morphologically poor and orthographically standardised language like Present-Day English.

This chapter presents the first systematic approach to annotating historical Welsh, ultimately aiming to provide an excellent starting point to build a fully annotated Welsh historical corpus (see also Meelen et al. 2017 and Meelen and Willis *forthc.*). This chapter discusses some of the challenges faced developing the methodology and annotating the first part of this corpus. Some of these challenges are specific to Welsh (or other Celtic languages), others inherent to working with historical corpora in general. Section 2 gives a brief overview of how the corpus was compiled and pre-processed. In Section 3, I discuss the procedure of part-of-speech (POS) tagging and of developing a tag set to annotate Welsh (and potentially other languages) with rich verbal and prepositional inflection. Sections 4 and 5 shed light on adding syntactic and information-structural features to the data so that it can be queried in various ways. In the final section, I demonstrate how this first annotated corpus of Middle Welsh prose can be extended and that this entails promising opportunities for future research.

2 Pre-processing

2.1 General philosophy and goals

Ideally, any corpus should be well-balanced in terms of text genre, length, origin etc. When working with historical data, however, the choices are often limited, and creating detailed annotation is extremely time-consuming. Therefore, this first annotated historical corpus of Middle Welsh focussed on the most commonly used editions² of the 11 native tales of the *Mabinogion* and some excerpts from the laws only. Future extension of the corpus will include alternative manuscript versions

² The editions used for this first annotated corpus are Williams (1951) for the *Four Branches*, Bromwich and Evans (1992) for *Culhwch and Olwen*, Thomson (1997) for *Gereint*, Thomson (1968) for *Owein*, Goetinck (1976) for *Peredur*, Roberts (1975) for *Lludd and Llefelys*, Williams (1908) for *Breuddwyd Maxen*, Richards (1948) for *Breudwyd Ronabwy*. For the new and complete

of each of these texts. In addition to that, the corpus will be extended to include more texts from different genres and registers such as the historical chronicles of the kings and princes and translated texts such as chronicles or Bible translations.

The main aim of this project was not to give a complete syntactic analysis or to provide a detailed parsed structure. The part-of-speech (POS) tags contain highly detailed morphological information (see Section 3), but the phrasal annotation is confined to a hierarchical chunk parse (see Section 4). In this way, the annotated corpus could remain theory-neutral. At the same time, queries for linear order and basic hierarchical phrase structure (e.g. to find noun phrases within prepositional phrases) are still possible. Finally, future enrichment of the chunk-parsed corpus is enabled, because of its flexible XML format (see Section 5).

Any controversial decisions were avoided as much as possible by backing off to a simpler form of annotation. For example, interjections were simply labelled as 'INTJ' instead of aiming to classify them further. Similarly, constructions that are changing over time were consistently annotated to facilitate future studies. A good example is the *sef*-construction³ in Welsh. The information-structural status of this construction changes from initial identificatory focus to plain predicate focus in the course of the Middle Welsh period (see Meelen 2016: 272–284). Since most texts are difficult to date exactly, throughout the corpus the specific tag SEF for any occurrence of this type of sentence was used (see also Section 3 on POS tagging). In this way, all sentences with the *sef*-construction can be easily found and examined in context.

2.2 Preparation and text formatting

There are various orthographical inconsistencies in the various versions of the texts of the *Mabinogion* (cf. Thomson 1986: xi), e.g. *yniuerod~y niuerod* 'numbers', *mywn~mewn* 'in', etc. For the present study, the texts were not extensively pre-processed through text normalisation or orthographical regularisation, because there was no stemmer available yet for Middle or Early Modern Welsh. A stemmer can automatically detect the root or stem of inflected and conjugated

Parsed Historical Corpus of the Welsh Language (PARSHCWL), we will create new editions so that all annotated texts can be deposited and made freely available to any scholars.

³ This construction originated from a reduced identificatory cleft construction *ys ef* 'it is it', but developed in the course of the Old and Middle Welsh period into an adverb *sef* meaning 'namely'.

forms, which is extremely useful when automatically processing a new text in a particular language. With a stemmer the task of POS tagging is easier because the various forms are reduced to stems + their respective inflection/conjugation. Developing a good stemmer or carrying out the procedure manually during the pre-processing stage is a tremendous task, however, which is why the detection of inflected forms was left to the POS-tagging stage, where the automatic classification algorithm could use specific features to disambiguate words regardless of their varying orthography (see Section 3).

In order to prepare the texts for annotation, a minimum amount of preparation is always necessary, however. First of all, the markup was stripped from the digitised texts, which were then saved as plain text files (.txt) so that they would be in the right input format for the POS tagger. Further pre-processing involved the insertion of sentence-final punctuation (if that was not present in the edition already) and the deletion of sentence-internal full stops. Finally, utterance boundaries in the form of <utt> were inserted semi-automatically (automatically after a full stop and manually if the full stop did not exist). Without utterance boundaries, the POS tagger is not able to assign morphosyntactic tags to all tokens. In addition, utterance boundaries are useful units for subsequent NLP tasks like chunking, chinking⁴ and full syntactic parses. The only punctuation marks that were removed were the full stops preceding and following numbers, e.g. ‘.12.’ was turned into ‘12’ to optimise automatic token recognition. Tokenisation (the isolation of words) was done automatically by the POS tagger on the basis of word spacing and full stops at the end of an utterance.

The Natural Language Toolkit (NLTK, www.nltk.org) regular expression chunk-parser (see Section 4) requires a list of words and tags. Therefore, the POS-tagged text files of the format token/TAG were converted to the right input format for the chunk-parser using a simple text conversion script designed especially for this purpose (see Meelen 2016: 326 for a sample of the code and further details of the chunking process).

2.3 Splitting and joining tokens

Meelen and Beekhuizen (2013) in their pilot study observed that the huge amount of orthographic variation in Middle Welsh complicates the POS-tagging task tremendously. The advantage of using a Memory-Based Tagger (MBT) is that this

⁴ Chunking is defining a set of words that should be grouped into phrases; chinking is defining the set of words that should be excluded from those phrases.

type of tagger could filter those out on the basis of the context most of the time (see Section 3). Splitting and joining certain tokens still had to be done, however. When two words are merged together without spaces in the orthography, e.g. *ymangor* (*yn* + *Bangor*) ‘in Bangor’, the resulting combination contains a preposition ‘in’ and place name ‘Bangor’, but only receives one POS tag. If words like these are not split, it is difficult to decide which morphosyntactic tag this should be (preposition or place name). Joint tags like ‘P-NPR’ (‘preposition + proper noun’) could in theory be created, but the larger the amount of POS tags, the more difficult it will be for the algorithm to classify words correctly. Also, a researcher interested in place names would need the opportunity to automatically extract just place names from the corpus and would not like to be confronted with extra work sifting through examples of place names combined with prepositions (in case of combined tags) or missing examples altogether (if only the preposition tag was used). Some tokens, however, were particularly challenging for the automated tagger, since very few generalisations could be made from the small training set (see Meelen and Beekhuizen 2013). To overcome some of those very specific orthographic challenges, the following combinations were automatically split using so-called regular expression replacements. Regular expressions are sequences of characters that define search patterns. They allow for more detailed searches (and therefore quicker semi-automatic replacements) than simple string searches, because they can include logical operators. In this way, fixed combinations with prepositions that cause nasal mutation like *yn* ‘in’, e.g. *ymwyt* > *y#*⁵ + *mwyt* ‘in food’, conjunctions combined with definite articles: *ar* > *a#* + *r* ‘and the’ and particles combined with pronouns, e.g. *ae* > *a#* + *e* ‘PCL PRO-A’ (particle + accusative pronoun).

3 POS tagging

A properly pre-processed version can be tagged automatically by a part-of-speech tagger. Although for this first attempt to create an annotated corpus pre-processing tasks were minimal, the tokenisation (splitting and combining words) alongside the insertion of sentence boundaries was enough to make a memory-based algorithm perform well. For Middle Welsh, there was no Part-of-Speech tagger available yet. I therefore generated an MBT that could subsequently be used to assign morphosyntactic tags to the Middle Welsh data. For this purpose,

⁵ The # sign was inserted to indicate word breaks so that researchers are still able to identify the original orthography.

decisions have to be made concerning the tag set, the list of morphosyntactic labels for each of the Middle Welsh words. A very detailed tag set facilitates more (and different types of) research. When working with a corpus of limited size, however, use of too many different tags leads to low frequencies and many hapaxes, which in turn complicates the automatic tagging task and yields degraded results. In this section I describe these challenges and furthermore offer some solutions that are not only useful for those working on Middle Welsh, but for anyone working with similarly complex historical data.

3.1 Establishing the morphosyntactic tag set

The rich morphology, the initial consonant mutations and the abundant orthographical variations of Middle Welsh pose significant problems for any automated task in Natural Language Processing. Extensive pre-processing including normalisation of spelling and mutations, for example, would simplify the POS tagging since the tagger would then be able to recognise a larger number of words. Apart from the fact that such regularisations are time-consuming, it entails a type of editorial intervention that has major implications for future research. When orthography is regularised, for example, editorial decisions have to be made concerning what this ‘regular’ spelling should be in the first place. It also does not allow for dialectal variation or differences in scribal practices that might give crucial insights in the linguistic history and geography at a given time. Another way to simplify the tagging task is to limit the tag set to a short list of broad morphosyntactic categories like ‘**VERB**’, ‘**NOUN**’, etc. However, this too limits the range of research opportunities tremendously. It is therefore worthwhile to develop an annotation scheme that gives as much morphosyntactic information as possible.

A commonly used annotation scheme adding morphosyntactic information to historical corpora is the ‘UPenn standard’ developed initially for Old and Middle English texts (see www.ling.upenn.edu/histcorpora). This annotation scheme, however, does not always provide enough information to answer certain research questions, mainly queries concerning agreement patterns and changes in information structure. To enable further research in these and other areas, I have used the standardised UPenn scheme, but extended the part-of-speech tag set where necessary. Starting from the already extended tag set used for the Icelandic corpus (cf. Wallenberg et al. 2011), I have examined the features of Middle Welsh grammar and systematically added extra features, i.e. more inflectional features such as person and number added to the root and tense/mood forms with a dash/hyphen.

3.1.1 Verbal tags

Verbal inflection in Welsh mainly occurs as a suffix to the verbal stem. Inflected verbs in the UPenn tag set are tagged VB. Past tense is indicated by the regular English past-tense ending in *-ed*, resulting in VBD. For Welsh, the VBD for the preterite tense was kept and in the same way tags for present (-P), future (-F) (only relevant for irregular verbs) and pluperfect (-G, for Welsh *gorberffaith* ‘pluperfect’), imperative (-I) and imperfect (-A, for Welsh *amherffaith* ‘imperfect’) etc. were added. Finally, a distinction was made between indicative (-I) or subjunctive (-S) mood for the present and imperfect tenses.⁶ This results in insightful systematic combinations like VBPI (present indicative), VBAI (imperfect indicative), VBG (pluperfect) etc. The same letters were systematically added to irregular verbs, resulting in for example DOPI (present indicative of the verb *gwneuthur* ‘to do’), GTI⁷ (imperative of the verb *cael* ‘to get’) or BEAS (imperfect subjunctive of the verb *bod* ‘to be’).

Apart from these more-detailed tense-aspect-mood markers, further information was added about the inflection to indicate person and number. Following standard glossing practices, person and number were represented as -1SG (first-person singular), -2PL (second-person plural) etc. Welsh has a further inflectional suffix for the ‘impersonal’ form of the verb that can be used in true impersonal contexts meaning ‘one’ or underspecified ‘they’; it is frequently translated into English as a passive. The number 4 was employed for this specific suffix and added to the verbal tags like the other personal endings, e.g. VBPI-4 (impersonal present indicative) or DOAI-4 (impersonal imperfect indicative of the verb *gwneuthur* ‘to do’).

3.1.2 Inflected and combined prepositions

Another feature of the grammar, specific to Welsh and other Celtic languages (but also seen in for example Semitic languages like Arabic or Hebrew), is inflected prepositions. Middle Welsh had a specific set of prepositions that could

⁶ Note that the ‘-I’ for “indicative” can only appear combined with ‘-P-’ (present) or ‘-A-’ (imperfect); the ‘-I’ for ‘imperative’ only appears directly after the core verbal tag (VB/GT/BE/DO). Since the imperative does not have different tenses this does not lead to any ambiguities in the annotation system.

⁷ The initial ‘HV-’ is for the auxiliary verb *have* in the UPenn standard annotation scheme. In Welsh, *cael* can mean ‘have, get’ with auxiliary functions as well, although it is not the exact equivalent of English *have*, which is why GT ‘get’ was chosen for Welsh *cael*.

be inflected for person, number and gender (the last of these in the third-person singular only and only ever with a pronominal object). There are also ‘uninflected’ prepositions in Welsh, but the inflected set includes very common prepositions like *i* ‘to’, *ar* ‘on’ and *yn* ‘in’. Following the verbal inflectional tags, Middle Welsh *iddi* ‘to her’ for example was tagged as P-3SGF ‘preposition third-person singular feminine’.

Some prepositions in Welsh could also be combined with other prepositions, e.g. *y dan* ‘under, below’ from *y* ‘to’ + *tan* ‘under’. These complex prepositions were tagged P + PX, so they could be recognised as separate, but also as combined prepositions. A further advantage of this is that the automatic tagger looking at the tags preceding and following the focus word, will not encounter the rare sequence of two prepositions. A disadvantage remains, of course, that the tag set is further extended and there are more homophonous forms that could render worse results if the complex preposition in question does not occur frequently in the training set.

Welsh also allows for some further types of complex prepositions: a combination of a preposition plus a grammaticalised noun. If the object of this type of preposition is a pronoun, it can appear in between the two prepositions as a possessive pronoun, e.g. *yn eu herbyn* ‘against/towards them’ from *yn* ‘in’ plus *eu* ‘their’ plus *erbyn* ‘opposition’. There are two possible ways to annotate constructions that are changing in historical corpora: we can annotate the original structure and form or the new construction as a whole. Since the exact date of grammaticalisation is often difficult to determine, it is not always easy to choose one or the other. As long as the construction is tagged consistently in one text (or one period of the historical corpus) and the annotation manual is clear about this, this should not be a problem. In that case, future researchers will always be able to find and, if necessary, change the annotation again. In this particular case of combined prepositions, a less conservative annotation scheme, disregarding the nominal origin of the construction yielding the tag sequence ‘P PRO-G PX’ (preposition – possessive pronoun – second part of combined preposition) was preferred to facilitate research into prepositional phrases.

3.1.3 Distinguishing different types of pronominal forms

Another part of grammar in which the tag set was extended significantly is pronominal forms. Since Welsh has various sets of pronouns for different (grammatical) contexts, a more fine-grained distinction here could enhance research not only in the pronominal domain, but also in Information Structure. Conjunctive pronouns like *ynteu* ‘he (then)’, for example, are used in contexts of

topic switch, meaning ‘but I’, ‘I, then,’ etc. Reduplicated pronouns like *tydi* ‘you’, on the other hand, are only used in focussed contexts. Separate tags for those are thus useful for finding the focus domain of sentences.

A further distinction in the pronominal domain was made between possessive pronouns and object pronouns. Since the infixed versions of these pronouns often exhibit the exact same form, a more fine-grained distinction in the tag set facilitates syntactic research here as well. Following the extensions of the tag set for the Icelandic parsed corpus (see Wallenberg et al. 2011), these pronominal tags receive case tags marked with a dash, for example, *fy* ‘my’ PRO-G (‘pronoun genitive’), or *e* ‘him’ PRO-A (‘pronoun accusative’).

3.1.4 Additional extensions of the tag set

Further extensions of the tag set include ADJQ for equative constructions, e.g. *cochet* ‘as red’ (from *coch* ‘red’ + equative *-et*) and ADJPL for plural adjectives, e.g. *gueisson ieueinc* ‘young servants’. More detailed tags like these are helpful to historical linguists and syntacticians looking at the structure and agreement patterns of noun phrases.

As described above, Welsh employs a wide range of particles. These too were tagged separately according to their function (e.g. PCL-QU ‘question particle’, PCL-FOC ‘focus particle’, PCL-NEG ‘negative particle’) to help distinguish different types of clauses. Aspectual particles like *yn* ‘progressive’ (PROGR) or *wedi* ‘perfective’ (PERF) were also distinguished from the homophonous predicative particles (PRED) and prepositions (P) respectively.

The verbal noun category characteristic of Celtic languages was tagged VN for regular verbs. Irregular verbs with verbal nouns that have specific functions in Welsh, e.g. *cael* ‘get’, also used for the passive, received specific verbal noun tags. The -N was added systematically to their base forms, e.g. GT- ‘have, get’ > GTN ‘verbal noun of the verb *cael* ‘to get’’. The verbal noun of the verb ‘to be’ was tagged as ‘BEN’, although it can also appear in this form in many other syntactic contexts, e.g. in complement clauses.

Finally, some additional lexical items with specific functions were tagged separately. An example of this is the above-mentioned petrified form *sef* (tagged SEF), which was used in earlier stages of the language to focus identificatory copular sentences. During the Middle Welsh period, it grammaticalised further until it became an adverbial element used in apposition to noun phrases meaning ‘that is’ (cf. Latin *id est* still used as the abbreviation *i.e.* in English).

One final problem that remained for Middle Welsh was the large amount of homophony alluded to in various cases above already. Because there is a large number of very short words in Middle Welsh that are spelled the same with a wide range of meanings, e.g. *a* which can mean ‘and’, ‘with’, preverbal particle, etc. this poses problems for automatic classification algorithms. The tagger, however, was often able to distinguish between up to five different possible meanings of, for example, Middle Welsh *y* ‘the, his, her, to, to his/her, in’ etc. on the basis of the preceding and following context.

3.2 Morphosyntactic annotation

With the MBT from TiMBL⁸ it is possible to generate a tagger for any language based on a training set consisting of a systematic token/TAG format with utterance boundaries at the end of every sentential unit. A memory-based tagger uses memory-based algorithms to disambiguate and classify words in a corpus. Unlike other types of taggers, memory-based taggers interweave processing and learning stages. Whenever a new language item + classifier is encountered in the training data, it leaves a memory trace that guides subsequent processing. This means that when a new instance is found and needs to be classified, a set of relevant instances is selected from memory with a number of useful features, and the new token is classified by analogy to that set. This therefore yields robust results especially when POS tagging languages with orthographical variation and rich morphology (see Zavrel and Daelemans 1997, and Daelemans and Van den Bosch 2005 for instructions, background and further functionalities).⁹

3.2.1 Automatic POS tagging

Once the Middle Welsh tagger is generated, the settings file of the tagger is used to assign POS tags to a new part of the corpus (presented as a tokenised text file). Based on the training set, the MBT divides the new text in need of annotation into ‘known’ and ‘unknown’ words. Depending on the exact

⁸ <https://languagemachines.github.io/mbt/>

⁹ For the new version of PARSHCWL, the results of this Memory-Based Tagger will be compared to a state-of-the-art BiLSTM-CNN-CRF tagger (see <https://github.com/achernodub/targer>) to see which of those yields better results and should form the basis of texts that need to be annotated and added to PARSHCWL in the future (see Meelen and Willis *forthc.*).

parameter settings and the features from the training set stored in memory, the tagger will then assign a tag to each word.

In Welsh the inflection appears as a suffix (on verbs or prepositions). When the tagger finds an unknown word like *ohonaf* ‘of me’, for example, it can compare the last three characters to known words with assigned tags in the training set. An example of this could be another inflected preposition, like *arnaf* ‘on me’ with the POS tag P-1SG (‘preposition + first person singular ending’). The exact same final characters (in combination with the other tags in the preceding and following context) lead the MBT to assign the same tag P-1SG to *ohonaf*, which would be the correct tag. Note that the context of inflected verbs and prepositions is often quite distinct (e.g. prepositions can appear after inflected verbs whereas inflected verbs do not). Since the automatic tagger is also sensitive to context, it will be able to distinguish verbs from prepositions and therefore verbs ending in *-naf*, e.g. *canaf* ‘I sing’, would be correctly tagged as first-person singular present verbs (VBPI-1SG) instead of inflected prepositions (P-1SG).¹⁰

Known words are easier if there are no homophones with different tags. If there are homophones, for example the above-mentioned Middle Welsh word *y*, the context in which they appear is crucial. In between an adverb (ADV) and an inflected verb (VB*/GT*/BE*/DO*), *y* is undoubtedly the preverbal particle following sentence-initial adjuncts, like in (1a). In front of verbal nouns, however, like at the end of (1b), *y* could be the preposition ‘to’ or a possessive pronoun (masculine, feminine or third-person plural), as earlier in (1b).

- (1) a. *Tranhoeth y deuthant y ’r llys.*
 next.day PTC come_{3PL,PRET} to DEF court
 ‘The next day they came to the court.’ (Bromwich and Evans 1992: line 595, [*Culhwch ac Olwen*])
- b. *a dyuot yn y uryt ac yn*
 and come_{VN} in 3SG_{MASC.POSS} mind and in
y uedwl uynet y hela
 3SG_{MASC.POSS} mind go_{VN} to hunt_{VN}
 ‘and he was minded to go and hunt’ (Williams 1951: lines 3–4 [*Pwyll Pendeuic Dyuet*])

¹⁰ Many thanks to an anonymous reviewer who suggested explaining this potentially ambiguous case.

The output file of the tagging process is a text file consisting of a word + TAG (shown in Figure 1) and an indication whether this word was known (signalled by a single forward slash “/”) or unknown (signalled by a double forward slash “//”) from the training set.

```
Kilyd//NPR mab/N Kyledon//NPR Wledic//NPR a/PCL uynnei/VBAI-
3SG wreic/N kynmwyd/ADJQ ac/P ef/PRO ./PUNC <utt>
Sef/SEF gwreic/N a/PCL uynnwys/VBD-3SG ,/PUNC Goleudyd//NPR
merch/N Anlawd//NPR Wledic//NPR ./PUNC <utt>
```

Figure 1: Fragment of the output of the POS tagger for the text *Culhwch and Olwen*.

MBT allows for different settings according to the features of the words themselves or the context in which they appear. In order to obtain the maximally reliable tags, a wide range of parameter settings was tried, varying those features. The optimal settings for Middle Welsh known (-p) and unknown (-P) words are the following (see Meelen 2016: 331–332 and Daelemans et al. 2010 for further details on these specific settings):

- p dfa
- P sssdFawchn
- M 200 -n 5 -% 5 -O +vS -F Columns-G K: -a 0 U: -a 0 -m M -k 17 -d IL

For Middle Welsh, the corrected gold standard of one text was subsequently used to annotate other texts of the *Mabinogion* and the laws automatically with greater accuracy. Each of those texts was in turn manually corrected as well.

3.2.2 POS tagging results

In order to estimate the quality of the POS tagger and obtain optimal parameter settings, I evaluated the manually annotated data with a ten-fold cross-validation, i.e. taking 90% of the data, training the model on that subset and then testing it on the other 10%, repeating this procedure for ten 90%–10% splits. Since the ten percent that the model is tested on is manually checked, we can see how often the model assigns the correct tag to a word, as well as obtain insightful statistics about the over- and under-generalisations of some tags. The above-mentioned settings gave the following results for the 59,000-word Middle Welsh corpus (see Meelen 2016: 40–44 for a full overview and further discussion of the results):

- Global Accuracy: 90.4%
- Global Accuracy known words: 93.3%
- Global Accuracy unknown words: 63.3%

To give a better insight in the success of the tagger, I calculated the Precision (percentage of system-provided tags that were correct), Recall (percentage of tags in the input that were correctly identified by the system) and F-score (weighted harmonic mean of Recall and Precision). We find high results for simple tags like CONJ ‘conjunction’ or N ‘noun’ that occur extremely often. As expected, Precision and Recall for tags occurring only once or twice is extremely low. These tags are often forms of verbs that occur very infrequently with irregular endings. Precision and Recall give more insight in the degree to which the model over- or under-generalises certain tags for the individual categories. The genitive (possessive) pronoun category (PRO-G), for instance, is correct about 90% of the cases where it is applied (90% Precision), but out of all actual possessive pronouns, only 65% is recognised (Recall of 65%). This is understandable, because the possessive pronoun usually consists of only one letter that is homophonous with the object infixed pronoun. The model thus under-generalised that category in particular. If 95% of the actual conjunctions on the other hand are recognised as such, while the item is only classified as a conjunction correctly in 90% of the cases, this category would be slightly over-generalised. As expected, the F-score for frequently occurring tags is considerably higher than that for tags and tokens occurring only once or twice in the corpus. The extremely fine-grained tag set with over 200 morphosyntactic labels (see the Appendix of Meelen 2016 for a full overview) can thus only reach a good Global Accuracy in a large corpus. This first corpus is not very large, which is why a Global Accuracy of over 90% is an acceptable result.

To conclude this section, Middle Welsh presents a good test case for POS tagging a historical corpus of a language with rich verbal and prepositional inflection and non-standardised orthography. The MBT showed robust results and flexibility with the highly variable orthography of minimally pre-processed Welsh texts. The parameter settings of the MBT software allow for focus on the context and the last 3 letters of unknown words. Since Middle Welsh verbal endings usually consist of 2/3-letter suffixes (reflecting tense, mood, aspect, person and number combined), it is not difficult for the tagger to predict the right form (e.g. *gwel-eis* ‘I saw’ as VBD-1SG denoting ‘first-person singular preterite’). Other parameter settings like an additional focus on the first 3 letters of the word proved to be less helpful for a language like Welsh with initial consonant mutation. This might, however, improve the results for languages with a strong prefixing preference, like for example Navajo (Young and Morgan 1980: 103, 107).

4 Annotating syntax and information structure

In order to facilitate syntactic queries, the above-described morphosyntactic annotation was employed to develop hierarchical phrase structure as well. A full parse would require a detailed Context-Free Grammar or Dependency Grammar. Developing this was beyond the scope of the present study, however. Instead, I modified the rule-based chunk-parser available in the Natural Language Toolkit (NLTK) in such a way that not only phrasal chunks but also some theory-neutral hierarchical structure could be added.

4.1 Designing the rule-based grammar

The NLTK rule-based chunk-parser is a regular expression parser: it systematically combines POS tags as defined in a grammar that allows regular expressions to create more (specific) options. Frequently used regular expressions include:

? $\Rightarrow$ for optional preceding items | $\Rightarrow$ ‘or’

The combination of words with their POS tags into phrases, e.g. Noun Phrases (NPs), Determiner Phrases (DPs) or Prepositional Phrases (PPs), is achieved with the following sample pattern of commands:

- NP: {<N|NPL|NPR>}
- DP: {<D><ADJ|ADJPL>?<NP>}
- PP: {<P><NP|DP>}

According to the above set of rules, a noun phrase (NP) can be formed of words with one of three different POS tags: a noun (N) *or* a plural noun (NPL) *or* a proper noun (NPR). Similarly, a DP, in this grammar, is formed by a determiner (D) followed by a noun phrase (NP) with an optional adjective (singular ADJ or plural ADJPL) in between. The order in which this rule-based grammar operates is important. The DP-rule above must follow the NP-rule to find the label <NP>. In this way single-layered hierarchical structures (NPs within DPs) were created. Similarly, a further layer could be created resulting in a PP containing a DP containing an NP, as long as they are called in the right order.

This is all straightforward in a language with extremely simple noun phrases and/or with a very limited amount of POS tags. Middle Welsh noun phrases, however, present some problems in this respect. First of all, some adjectives either follow or precede the noun they modify, with different meanings in either of the two positions. In addition to this, possessive pronouns and quantifiers can be part of the noun phrase as well. Furthermore, demonstratives

must follow the noun (and its modifying adjectives) and they are also obligatorily accompanied by the definite article preceding the noun phrase, as in (2). Finally, phrases with numerals in Welsh come in many shapes and forms, as (3) shows. Welsh numerals above ten can be split to occur before and after the noun phrase. In addition to that, phrases with numerals can also employ the preposition *o* ‘of’.

- (2) a. *y cathod mawr*
 DEF cats big
 ‘the big cats’
 b. *yr hen lyfr mawr hwn*
 DEF old book big PROX
 ‘this big old book’
- (3) a. *tair o ferched*
 three_{FEM} of girls
 ‘three girls’
 b. *tri o bobl eraill/newydd*
 three_{MASC} of people other_{PL}/new
 ‘three other/new people’
 c. *dau hen lyfr lyfr*
 two_{MASC} old book book
 ‘two old books’
 d. *un mlynedd ar ddeg*
 one year on ten
 ‘eleven years’

Complex noun phrases can also consist of two juxtaposed nouns, as in (4). In these constructions, the definite article only appears before the second noun, but the whole construction is definite.

- (4) a. *dyn y siop*
 man DEF shop
 ‘the man of the shop’
 b. *pob yn ail fis*
 every PRED second month
 ‘every other month’
 c. *yr holl broblemau*
 DEF all problems
 ‘all the problems’

The above types of complex noun phrases require a very detailed rule-based grammar that includes all possible phrases, including some phrases with special labels to facilitate further syntactic queries, e.g. phrases with verbal nouns (which can function as infinitives or nouns). Such a rule-based grammar explicitly looks for sequences of, for example, numerals + adjectives + nouns such as *dau hen lyfr* ‘two old books’ in (3c). Similarly, when put in the right order, complex noun phrases with quantifiers can be combined with explicit searches for the sequence determiner + quantifier + noun (D>Q>N). The full rule-based grammar I designed can be found in Meelen (2016: Appendix). This is a flexible template that can easily be extended and adapted to achieve better results when more texts are added.

4.2 Manual correction

No automatic NLP task is 100% correct. The rule-based chunk-parsers performs very well with simple matrix clauses, but subordinate clauses and some complex DPs in particular need some correction. I manually corrected the entire corpus using CesaX. CesaX is a special software package developed by Erwin Komen to facilitate corpus-linguistic research (cf. Komen 2011). Another useful feature of CesaX is that it can automatically convert the chunk-parsed .psd-files to XML-files with a simplified TEI P5 header. These files can then be queried using CorpusSearch¹¹ or the XML-based XQuery language. Manual correction in CesaX is quick and easy, because of its graphic representation of the tree structures. Alternatively, the bracket representation shown below, can also be edited manually with any text editor if needed:

```
(S (DP (NP (N taryan))(ADJP (ADJ eur))(NP (N grwydyr)))
(VP (PCL a)(VBD-3PL dodasant))
(PP (P dan)(DP (PRO-G y)(NP (N penn)))) (, .))
```

Figure 2: Sample bracket representation.

¹¹ CorpusSearch is a query language that finds syntactic structures in a corpus of annotated sentence trees. It can be used as a development tool for building the corpus or as a research tool to find and collate results in a corpus (see Section 5.1).

The above output from the automatic chunk parser reflects the following example:

- (5) *Taryan eur grwydyr a dodassant dan y penn*
 shield gold enamelled PTC put_{3PL,PRET} under 3SG_{MASC,POSS} head
 They placed a gold enamelled shield under his head.’ (Williams 1908:
 lines 18–19 [*Breuddwyd Maxen*])

4.3 Annotating information structure

Information-structural features were added semi-automatically. With CorpusStudio software (see Komen 2009 and Section 5), various features can be automatically added. Information for these features can be derived from the detailed POS tags of the specific words, from the phrasal structure and/or from the context in which it occurs. For example, since personal pronominal subjects usually convey ‘old’ information, with some simple XQuery commands the referential status of these subject pronouns can be automatically labelled ‘Old’. Other specific features of the clause such as the tense, aspect or mood of the verb or the person-number inflection can be derived from the detailed set of POS tags in the same way.

Further information-structural notions such as topic or focus are not as easy to detect automatically. However, if special focus words, pronouns or particles are used, these were labelled as such by the detailed POS tagger and therefore the focus domain or articulation of the sentence can automatically be annotated accordingly. In addition to this, constituent focus in Middle Welsh could be indicated by a (reduced) cleft and a verb with default third-person singular inflection. Pronominal subjects in the first or second person or plural full DPs can be automatically detected as well. When it comes to labelling the exact type of topic (e.g. familiar, aboutness or contrastive) or focus, manual annotation is still required.

All additional features (including the information-structural ones discussed here) are added at the matrix clause level. In practice, this means a list of features with automatically derived values (by querying the POS tags) and open values (to be adjusted manually) is available for every matrix clause. These features include:¹²

¹² These features are chosen because combined they cover all relevant information-structural notions. See Meelen (2016: Chapter 2) for further motivation with detailed examples from Middle Welsh and other languages.

- Focus articulation, e.g. constituent focus
- Focus particle/word, e.g. *hefyd* ‘also’
- Point of departure, e.g. temporal clause ‘at that moment . . .’
- Information flow, e.g. unmarked
- Referential state subject, e.g. old information
- Referential state object, e.g. new information
- Diathesis, e.g. impersonal verb
- Tense/aspect, e.g. imperfect
- Mood, e.g. subjunctive
- Semantic roles (in order), e.g. agent-patient
- Animacy and definiteness subject, e.g. definite-animate
- Animacy and definiteness object, e.g. indefinite-inanimate

5 A brief note on possibilities to query the data

There are various online tools available for corpus research, e.g. the search interface for the British National Corpus.¹³ Search interfaces provide easy access to the data, because no prior knowledge of specific search algorithms is necessary to get any results. The types of searches are often limited to the level of individual words or simple part-of-speech labels, however, which is not sufficient for syntactic or more detailed linguistic analyses. If we want to gain a deeper insight in our linguistic data, we need a more thorough way of searching for the right information.

5.1 Extended querying with CorpusStudio and CesaX

For many historical syntacticians, CorpusSearch¹⁴ is a useful application that can retrieve the detailed linguistic data relevant to them. It enables queries in the treebank or labelled bracketing format (the .psd format described above). With CorpusSearch, these first POS-tagged and chunk-parsed files can thus be easily queried and, for example, compared to data from historical corpora of other languages. It should be noted, however, that this first partial corpus only contains an extended shallow parse without syntactic empty categories yet.

¹³ www.natcorp.ox.ac.uk/

¹⁴ <http://corpussearch.sourceforge.net/>

These will, however, be added to PARSHCWL in the future (see Meelen and Willis *forthc.*).

Another way to retrieve detailed syntactic information is by converting the (parsed) files to XML format and query them with the usual search function for XML-databases: XQuery and XPath. Erwin Komen developed a wrapper around CorpusSearch2 (Randall, Taylor and Kroch 2005) and XQuery to facilitate these searches: CorpusStudio (Komen 2009). CorpusStudio not only simplifies the task of formulating search queries, it also provides easy ways to organise them along with the corpus data and research logs documenting your goals, subqueries, definition files and any emendations while gathering the right data.

5.2 Textual markup

For the metadata markup, I chose a simplified version of the TEI P5 (TEI Consortium 2009) header that is suitable for philological data, translations and linguistic annotation in XML format. This simplified TEI P5 header was selected because the full header with all its details was unnecessary and inefficient to work with. In addition, parsed files can be converted to this simplified TEI P5 header automatically by the CesaX software (see Section 4 and Komen 2011). Any information about the philological background of the text can be stored in this header and easily retrieved for future online usage. In the textual markup, any changes to the annotation can be indicated as well to trace the history of the annotated text and corpus as a whole. Finally, it would ultimately be possible to combine different versions of the texts (i.e. diplomatic and critical editions) into one XML file to make sure invaluable philological information is not lost. Its systematic but flexible nature would allow future conversion to JSON and RDF format as well.

6 Conclusion

This chapter presents the first steps towards the creation of a fully annotated corpus of historical Welsh. The above description of the proposed procedure is meant as a blueprint for the development of a fully parsed historical Welsh treebank (PARSHCWL) in the future (see Meelen and Willis *forthc.*). I described how a combination of minimal pre-processing, a systematic extension of current tag sets for historical corpora and a hierarchical way of chunk-parsing can yield important information needed to address questions about the syntax and

information structure of the Welsh language that have hitherto been unanswered. At the current stage, the annotation of the corpus was done in such a way as to optimise the search queries specific to the change of word order in the early Middle Welsh period (see Meelen 2016). However, the same annotated partial corpus of Middle Welsh was also already used for studies on adjectival agreement in native and translated Welsh prose (see Meelen and Nurmio 2020; Parina and Poppe *forthc.*).

The flexible XML-based nature (compatible with the .psd file structure) of the corpus means that any further philological or linguistic annotation can be added at a later stage as well. At various stages in the process of creating the corpus, manual correction was necessary. For this pilot, there was only one annotator available to do the manual correction of the limited pre-processing and more-detailed POS tagging and parsing tasks. Therefore, checking cross-annotator agreement, which is needed to verify the results, was not an option. In future, when making the annotated files accessible for everyone online, a final check will be done to filter out any possible mistakes and/or inconsistencies.

This chapter presents a good test case for annotating a partial historical corpus of a language with rich verbal and prepositional inflection. The main challenges in building annotated corpora like these lie in the availability of good digitised diplomatic or critical text editions. Further collaboration with scholars specialised in the philological background producing these editions can help linguists to make the right decisions, both in terms of selecting the right texts and editions for the corpus, but also in pre-processing and tokenisation in particular. More elaborate pre-processing of the texts, including the development of a good stemmer to do normalisation etc. and expanding the training set takes time, but will yield better results for the automatic POS tagging and parsing tasks in the end. A relatively large set of over >200 morpho-syntactic tags was developed and presented here, because those details give important new research opportunities for both Welsh philologists and linguists. A standardised way of expanding the tag set for rich inflectional languages is called for and the proposed extensions outlined in Section 3 above aims to be a good starting point for future extension and refinements when more texts are added to the corpus (see Meelen and Willis *forthc.* for this and subsequent steps towards building a fully parsed Welsh treebank).