JS Journal of Historical Syntax
Volume 6, Article 5: 1–32, 2022

TOWARDS A HISTORICAL TREEBANK OF MIDDLE AND MODERN WELSH: SYNTACTIC PARSING*

Marieke Meelen
University of Cambridge

David Willis
University of Oxford

Abstract This article examines various issues involved in constructing a parsed Penn-style representative historical corpus of Middle and Modern Welsh. Specifically, it focuses on what structures to adopt for constituency-based structural descriptions in three case studies: (i) whether to adopt relatively more or less hierarchical structures at the phrasal level and above; (ii) how to deal with complex prepositional phrases, typically containing a grammaticalizing or grammaticalized noun as one of their elements; and (iii) how to deal with coordination of main clauses and omission of elements shared between clauses. In each case, we see how conventions need to be adopted that facilitate maximal ease of searching for potential users of the corpus; that are robust across many centuries of language change; and that permit efficient and consistent parsing by a team of annotators.

1 Introduction

This article reviews aspects of the parsing system for the Parsed Historical Corpus of the Welsh Language (PARSHCWL). The goal of PARSHCWL is to create a fully parsed representative corpus of historical Welsh texts from the medieval period to the twentieth century to assist linguistics researchers with the task of extracting relevant syntactic structures of interest across the attested history of the language. In doing so, it builds on the existing His-

* We are grateful to the Arts and Humanities Research Council for funding the UK portion of the AHRC–DFG UK–German collaborative research project in the humanities 'The history of pronominal subjects in the languages of northern Europe' (award no. AH/V00347X/1). For insightful comments and suggestions, our thanks also go to three *JHS* anonymous reviewers and to George Walkden, Anne Breitbarth, Patrick Sims-Williams and to the workshop audience. Errors and shortcomings remain our own.

torical Corpus of the Welsh Language 1500–1850 (HCWL) (Willis & Mittendorf 2004), a representative but linguistically unannotated corpus of around 420,000 words from some 30 texts. PARSHCWL will incorporate most of the prose parts of HCWL, additionally expanding its time depth to cover the medieval period, in so doing creating a corpus of 100,000 words for the seven centuries from 1300 to 2000 plus a small representative set of the much sparser pre-1300 material. It aims to include a balanced range of historical, native narrative, translated narrative, religious and law texts from the rich attestation of medieval Welsh, and introduces extracts from novels, personal letters, drama, biography, Bible translations and journalism to the extent that these are available in later periods. PARSHCWL forms part of the Penn family of parsed historical corpora, which, in addition to a range of English corpora, such as the York–Toronto–Helsinki Parsed Corpus of Old English Prose (YCOE) and the Penn–Helsinki Parsed Corpus of Middle English (PPCME2), includes historical French, Portuguese, Irish, Old Saxon and Icelandic corpora. As with these corpora, each text is parsed at two levels: first, a Part-of-Speech (POS) tag is assigned to each individual word, and then a phrase-structure description (PSD) is created for each phrase. In order to facilitate parallel use of multiple corpora in the family, it is natural for PARSHCWL to start out from conventions common to those corpora at both of these levels. Nevertheless, extensions and deviations from those conventions are necessitated by properties of each individual language, and Welsh is no exception.

This article deals with three examples of the sorts of issues that arise in producing conventions for the phrase-structure descriptions associated with each clause of the corpus. These are intended as representative of the sorts of issues that arise repeatedly in constructing a parsed corpus of this nature. Specifically, we will begin by looking at a very broad question, namely how to represent clause structure, before moving on to two specific example areas, namely the representation of complex prepositions and the representation of clausal coordination. In each case, we aim to set out the possibilities from which a final choice had to be made, and the logic behind the final decision. Before looking at these case studies, some broad familiarity with the preprocessing and Part-of-Speech conventions of the corpus is necessary. These are briefly set out in the following paragraphs, before we introduce basic features of the phrase-structure annotation, and then turn to the main case studies at hand.

1.1 Background to preprocessing and Part-of-Speech tagging conventions

We begin by summarizing aspects of the preprocessing procedure and part-of-speech tagging conventions as necessary background both to the corpus

and to the main focus of this article, namely decisions on parsing syntactic structure. We will not motivate these decisions in detail, since they are discussed at greater length elsewhere; see Meelen & Willis (2021). Preprocessing of texts (for details, see below) was carried out in a semi-supervised fashion through a combination of regular-expression replacement rules for frequent splits and to resolve other tokenization issues alongside manual correction. Part-of-Speech tagging was done using a Memory-Based Tagger developed by Meelen (2016). Since there is very little training data available for Middle Welsh, a Memory-Based Tagger (Daelemans, Zavrel, Van den Bosch & Van der Sloot 2010) still yields better results than neural-network based taggers such as TARGER (Chernodub, Oliynyk, Heidenreich, Bondarenko, Hagen, Biemann & Panchenko 2019). A unique ID was assigned to each sentence automatically using a similar custom-made Python script after manual correction of word and sentence segmentation and POS tags.

Actual historical texts present practical difficulties concerning word divisions, textual errors and emendations. These issues necessitate a preprocessing stage, during which certain changes to the original texts as presented in the corpus files are made in order to facilitate parsing and information-structural annotation. All changes are documented within the corpus using one of four markers. This is a more elaborate system than the Penn Parsed English corpora, which use `$` for splits and emendations and avoid further such changes by adopting compound tags joined by a `+`-sign (e.g. `ADJ+NS` for an adjective and singular noun written as a single word). Compound tags are rejected for PARSHCWL because taggers have difficulty in assigning them accurately and they complicate the task of the user in designing queries.

Where what was written as a single word had to be split for the purposes of tagging, this is marked using `#`; thus, Middle Welsh may write *yr* 'to the' as one word, which is split for tagging as `y#/P r/D` (with POS tags `P` = preposition and `D` = determiner, separated from the original text by a forward slash).[1]

Where what was written as two words has had to be joined together for tagging, this is marked using `!`. For instance, in Middle Welsh, the inflected preposition *o honaf* 'of/from me' is sometimes written as two words, but is then tagged as `o!honaf/P-1SG`.

Sometimes, splitting occurs where a single character 'belongs' to two words. In these cases, that character is assigned to one of the words, and the 'gap' in the other word is marked using `+`. Thus, a verb and a following subject pronoun are sometimes written together in Middle Welsh, for in-

1 Penn-style historical corpora are inconsistent on whether tags are introduced by a forward slash or an underscore. We have followed the Penn Parsed Corpus of Middle English in using a forward slash.

stance, *gweleisti* instead of *gweleist ti* 'you (sg.) saw'. This would be tagged as `gweleis+/VBD-2SG ti/PRO`, with `+` indicating that an orthographic word has been split up for annotation and also indicating the loss of the final <t> character of the verb to the pronoun. In some cases, this means that `+` may simply indicate a word that has disappeared into neighbouring words, as with Middle Welsh *y*, which may represent *y* 'to' + *y* 'his', tagged therefore as `y/P +/PRO-G`, with the form of the second word represented only as `+`.

Finally, textual emendations, from whatever source, are indicated using an asterisk `*` to alert users to possible philological issues, such as editorial emendations or additions. For instance, early in the medieval *Pwyll* tale, we encounter the word `*chyfuarchaf/VBPI-1SG` 'I greet', POS-tagged as first-person singular present indicative verb. The asterisk indicates a philological issue: the word actually appears in the manuscript as *fuarchaf*, a form that makes no sense and is presumably a scribal error, corrected in the corpus. This may need to be investigated by the corpus user, depending on their research questions, but cannot be encoded in full in the corpus.

The POS tagset is based loosely on the Icelandic (IcePaHC) tagset (Wallenberg et al. 2011), although we have not hesitated to depart from that tagset where we thought it necessary. Nouns are marked for number (`N` or `NPL`), and proper nouns (`NPR`) are distinguished from common nouns.[2] Welsh nouns lack case morphology at all periods of the corpus, so case annotations for nouns are not used. Some pronouns do show case distinctions, in Middle as well as in Modern Literary Welsh; hence, pronouns may be marked for accusative (`PRO-A`) or genitive (`PRO-G`) case, or may be left unmarked for case (`PRO`). Verbs are divided into general lexical verbs (`VB`), *bod* 'be' (`BE`), *cael*/*caffael* 'get' (`GT`) and *gwneuthur* 'do' (`DO`). To these tags are suffixed tense, mood, person and number specifications; for instance, *gwelais* 'I saw' is tagged as `VBD-1SG` (where `D` = preterite), to be understood as first-person singular preterite lexical verb, and *bydd* 'it will be' (`BEF-3SG`) (where `F` = future), to be understood as third-person singular future of 'be'. Prepositions are marked for person, number and gender, for instance, *ohoni* 'of/from her' (`P-3SGF`), third-person singular feminine preposition.

Some additional tags, not found in the other Penn-style historical corpora, are needed to accommodate specific aspects of Welsh grammar. Most notable among these are `VN` 'verbal noun', `SEF` for the identificatory focus marker *sef*, and a wide range of particles, for instance, `PCL` for preverbal particles *a*, *y*(*r*), *mi* and *fe*, `PRED` for the predicate marker *yn* (used before predicate adjectives

2 Use of `NPL` departs from the IcePaHC corpus, which uses `NS` for plural common nouns; `NPL` was felt to be clearer and, unlike `NS`, not at risk of being misunderstood to mean 'singular noun'; see Meelen & Willis (2021) for fuller discussion of tagset decisions.

and nominals), PCL-NEG for negative particles, PCL-QU for question particles, PCL-FOC for focus particles, PROGR for the progressive aspect marker *yn*, PERF for the perfective aspect marker *wedi*, and PPCL for the presentative particles *llyma, dyma* (cf. French *voici*, *voilà*).

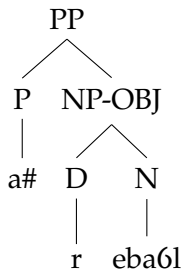
1.2 Basic parsing conventions

For this paper we focus on constituency-based phrase-structure annotation following parsing conventions of the Penn-style historical corpora (Santorini 2022). To facilitate adding phrase structure from scratch, we use the NLTK chunkparser, extended with hierarchical phrase-structure rules for Middle Welsh specifically, building on Meelen (2016). The syntactic annotation itself largely follows the Penn parsing guidelines, but there are some differences. Each full clause is marked as CP (complementizer phrase), with different clause types distinguished by suffixes: CP-MAT for matrix main clauses, CP-SUB for subordinate clauses, CP-REL for relative clauses and so on. This differs from Penn-style corpora, which use IP (inflectional phrase) as the clause marker. This choice is justified by the verb-second nature of Middle Welsh. Middle Welsh main clauses typically have the pattern topic phrase — preverbal particle — verb. Adopting a CP-layer allows the topic to be clearly identified as the element that lies within CP but outside IP. While it is true that parsed corpora have been developed for other V2-languages such as Old English, these languages lack a preverbal particle that clearly marks the division between the topic and other material, and they were developed at a time when avoidance of hierarchical structure was a more important criterion given limitations of earlier searching software.

Within clauses, phrases are identified for adjectives, adverbs, nouns and prepositions. These items always project their respective phrases ADJP, ADVP, NP, PP, even if the phrase only contains a single word. The internal structure of such phrases is relatively uncontroversial, and internal constituency can therefore be marked up. An example of a simple prepositional phrase containing a simple noun phrase is given below:

(1) *ar* *eba6l*
with.the foal
'with the foal' (RhG, Peniarth 4, folio 8v, 31.27, *Pwyll*)

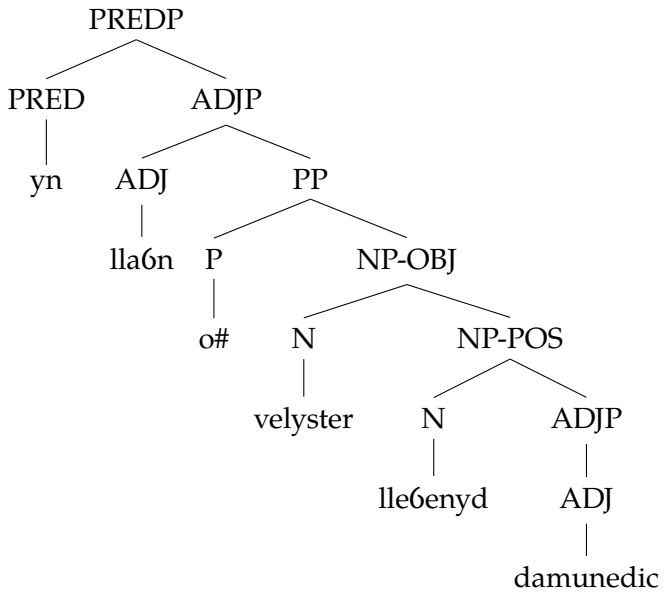
```
[PP [P a#]
    [NP-OBJ [D r] [N eba6l]]]
```



More elaborate examples, including nouns embedded inside other noun phrases, and nouns modified by an adjective and a prepositional phrase within another prepositional phrase, are given in (2) and (3).[3]

(2) *yn lla6n o velyster lle6enyd damunedic*
PRED full of sweetness joy longed.for
'full of the sweetness of longed-for joy' (RhG, Jesus 119, folio 111v, line 16, *Ystoria Adrian ac Ipotis*)

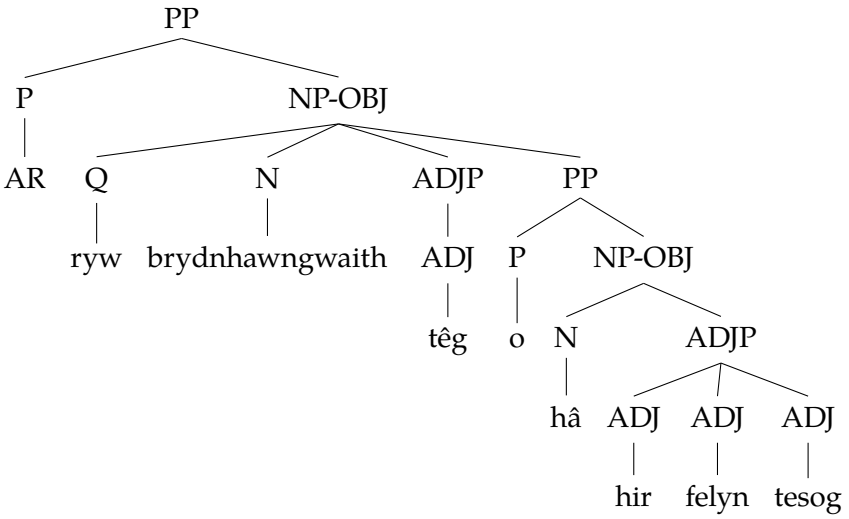
```
[PREDP  [PRED yn]
        [ADJP [ADJ lla6n]
              [PP [P o#]
                  [NP-OBJ [N velyster]
                          [NP-POS [N lle6enyd]
                                  [ADJP [ADJ damunedic]]]]]]]
```



3 For simplicity, we adopt a flat internal structure for the string of three adjectives in (3).

(3) *AR ryw brydnhawngwaith têg o hâ hir felyn tesog*
on some afternoon fair of summer long golden sultry
'on some fair afternoon of a long, golden, sultry summer' (HCWL, *Gweledigaetheu'r bardd cwsc*, p. 5, 1703)

```
[PP [P AR]
    [NP-OBJ [Q ryw]
            [N brydnhawngwaith]
            [ADJP [ADJ têg]]
            [PP [P o]
                [NP-OBJ [N hâ]
                        [ADJP [ADJ hir]
                              [ADJ felyn]
                              [ADJ tesog]]]]]]
```



Grammatical functions are indicated on noun phrases: `NP-SBJ` for subjects, `NP-OBJ` for objects, `NP-LOC` for locatives including directionals (`[NP-LOC adref]` '(to) home'), `NP-TMP` for noun phrases used as temporal adverbs (`[NP-TMP dyd Ieu]` '(on) Thursday'), `NP-PRD` for nominal predicates (`daeth yn [NP-PRD frenin]` 'became king'), `NP-POS` for possessives (`coron [NP-POS brenin]` 'a king's crown'), `NP-MSR` for measure phrases (`[NP-MSR filltir]` '(walked) a mile'), `NP-VOC` for vocatives (`[NP-VOC a frenin]` 'king!'), `NP-PRT` for apposition and parentheticals (`Dafydd [NP-PRT frenin]` 'King David'), `NP-RSP` for respect nominals (`bachgen du [NP-RSP ei wallt]` 'black-haired

boy (lit. boy black his hair)') and NP-ADT for other adjuncts. Features of Welsh syntax motivate the addition of NP-PRD and NP-RSP here, while some tags used in other Penn-style historical corpora were judged to be superfluous for Welsh (e.g. NP-SPD for secondary predicates, which can be incorporated into NP-PRD used for all predicate nominals). In order for such functions to be marked explicitly, subjects, objects and other nominals are always placed within a noun phrase, even if they are pronominal and even if they contain only a single unmodified word.

This introduction has outlined basic conventions at the clause level and in the internal structure of individual phrases. We now turn to look at the three individual case studies. We begin (section 2) with a major question raised by the discussion so far, namely how such elements fit into the broader picture of clause structure in the corpus. Once we have dealt with this question, we turn to a question of phrase-internal structure, namely the case of complex prepositons (section 3), before finally turning to a question which concerns inter-clausal relations, namely sentential coordination (section 4).

2 Flat vs. hierarchical structures

Every constituency-based treebank needs to address the issue of how to represent hierarchical structures. The default convention in the Penn historical corpora is to have fairly flat structures. The goal of annotation is not to provide a theoretical analysis, but a structural description that will be useful for searching (Santorini 2022: General introduction: Philosophy and goals). Thus, it does not matter if these structures deviate from the highly hierarchical structure and binary branching adopted in the theoretical frameworks within which analyses of these corpora have often been embedded, namely the Principles and Parameters and Minimalist frameworks. Adoption of relatively flat structures had several practical advantages. First, the original version of the CorpusSearch software (Randall, Taylor & Kroch 2005) that was used to query the earliest historical English corpora was faster and more efficient with flatter structures. Flatter structures furthermore avoid some potentially difficult choices among different options in the structure of phrases, especially where there is no scholarly agreement. Avoiding additional layers furthermore aids automatic parsing as there is less room for error. Finally, avoiding detailed hierarchical structures is advantageous as it makes the trees more theory-neutral and, as such, the treebank will be of use to more researchers using different frameworks. This section addresses the question of flat vs. hierarchical structures for PARSHCWL for verb phrases (VPs) with finite and nonfinite verbs.

2.1 VP layers for finite verbs

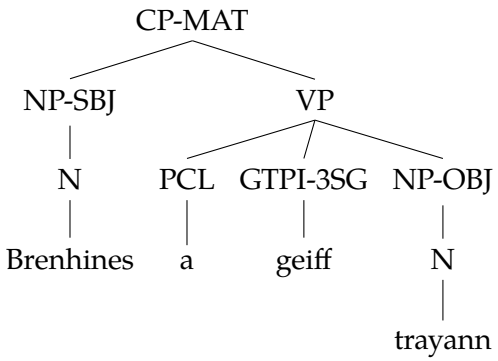
The Penn historical English corpora feature sentences with `IP-MAT` as the label for the top node of regular declarative matrix clauses. Furthermore, no clauses contain phrases labelled `VP`; verb phrases do not exist on any level, because decisions to keep structures relatively flat were taken early on. It is no longer necessary to keep structures flat for querying purposes: although speed can still play a small role, it is now possible to query any hierarchical structures just as easily as flat structures. The question then comes down to what would be theory-neutral, but at the same time remain insightful for the language under investigation.

Since Welsh has preverbal and aspectual particles and a wide variety of verbal-noun constructions with clitic pronouns, leaving out any form of hierarchy in the verb phrase may not be the best decision, as it makes the position and status of the preverbal particles and clitic pronouns less clear. Consider a typical Middle Welsh subject-initial verb-second clause such as (4): should the finite verb *geiff* 'receives' form a VP with the preverbal particle *a* and the postverbal direct object *trayann*?

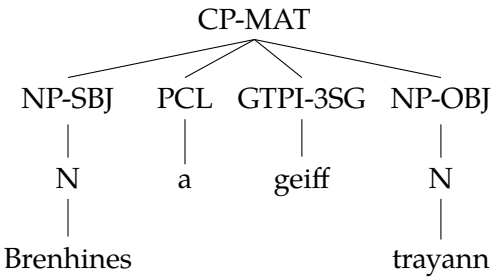
(4) *Brenhines* *a* *geiff* *trayann* [...]
queen PRT get.PRS.3SG third
'A queen receives a third [...]' (MCH, p. 1, clause 11, late 15th *c.*)
(subject V2)

Two possible structures for example (4) are:

(5)
```
[CP-MAT [NP-SBJ [N Brenhines]]
        [VP [PCL a]
            [GTPI-3SG geiff]
            [NP-OBJ [N trayann]]]]
```



(6) `[CP-MAT [NP-SBJ [N Brenhines]]`
`[PCL a]`
`[GTPI-3SG geiff]`
`[NP-OBJ [N trayann]]]`

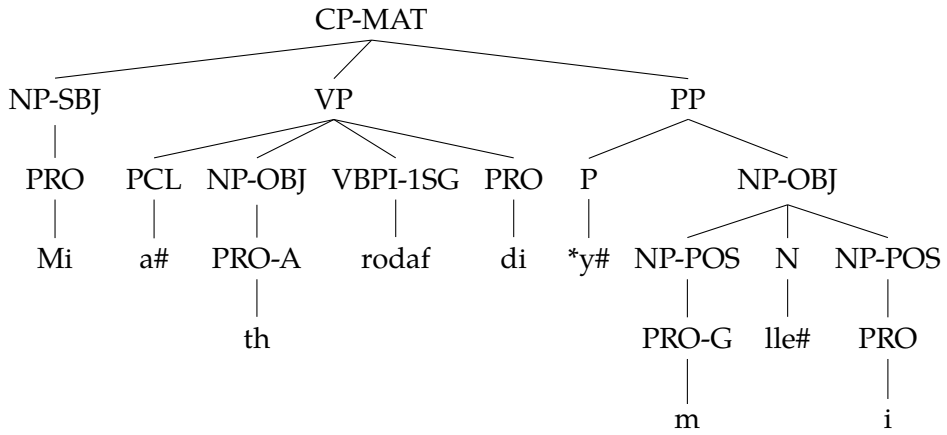


When the direct object is a pronoun, Middle Welsh uses an additional infixed accusative pronominal clitic between the preverbal particle and the inflected verb, as in (7).

(7) *Mi a# th rodaf di *y# m lle# i.*
I PRT you put.PRS.1SG you in 1SG place 1SG
'I will put you in my place.' (PKM 3.8, 14th *c.*) (V2 with infixed object)

Structurally, pronominal object clitics such as *'th* in (7) look similar to (although they are not the same as) the possessive genitive clitics found in the noun phrase (NP) *'m lle i* 'my place' and as the object of verbal nouns. If those are contained within an NP, then the object clitics (and the preverbal particles on which they are phonologically dependent) could be contained within a VP in a similar way, as illustrated by the following structure:

(8) `[CP-MAT [NP-SBJ [PRO Mi]]`
`[VP [PCL a#]`
`[NP-OBJ [PRO-A th]]`
`[VBPI-1SG rodaf]`
`[PRO di]]`
`[PP [P *y#]`
`[NP-OBJ [NP-POS [PRO-G m]]`
`[N lle#]`
`[NP-POS [PRO i]]]]]`



The pronominal object (accusative) and possessive (genitive) clitics, however, are by no means identical in form or structure (see Borsley, Tallerman & Willis 2007: 319–326). Furthermore, since the preverbal particle is a kind of clause-type marker, it is not an obvious candidate to be placed inside a VP.[4] Although the preverbal particle could under some analyses occupy the same structural position as the inflected verb (through head movement of the verb as incorporation into the complementizer C head; see Meelen 2016), this analysis is not necessarily the only option, and lower positions in more finely articulated IP and CP layers remain viable alternatives to host the verb (Willis 1998: 68–71). To adhere to the principle of avoiding an explicit choice between different structural analyses, making the treebank more theory-neutral, it is best to avoid use of a VP layer for inflected verb phrases. The structure lacking VP, in (6), is therefore preferred over those with a VP-layer in (5) and (8).

Another reason to reject VP layers in these cases comes from the V2 nature of Middle Welsh. Although the structure with the VP layers above reflects the close cohesion between the preverbal particle *a* and the infixed object and verb, as well as the fact that both the nominal and pronominal objects are clearly internal arguments in that same VP, adding this VP layer would be a real challenge for an automated Middle Welsh parser. Because of the V2 nature of the language, object-initial clauses and null subjects, both of which are illustrated in (9), are found frequently.

(9) *Ereill* *a* *dilea6d.*
others PRT destroy.PST.3SG
'Others he destroyed.' (MCH, p. 1, clause 7, late 15th *c.*) (object V2)

4 Note, however, that *a geiff* can correspond to relative *yssyd* 'which is', which is a single word. This would be an argument for having VP include the particle *a*, in order to allow a parallel treatment of *a geiff* and *yssyd*.

Thus an initial nominal element may be a subject or an object (or indeed the fronted object of a preposition, a predicate nominal, or a possessor). Often the distinction between subject- and object-initial orders becomes clear from context, but contextual clues are notoriously difficult for automatic parsers and, even with context, some ambiguities remain. Given the advantages and disadvantages of both options, it appears to be preferable to avoid using VP layers for finite verbs in PARSHCWL.

2.2 VP layers for nonfinite verbs

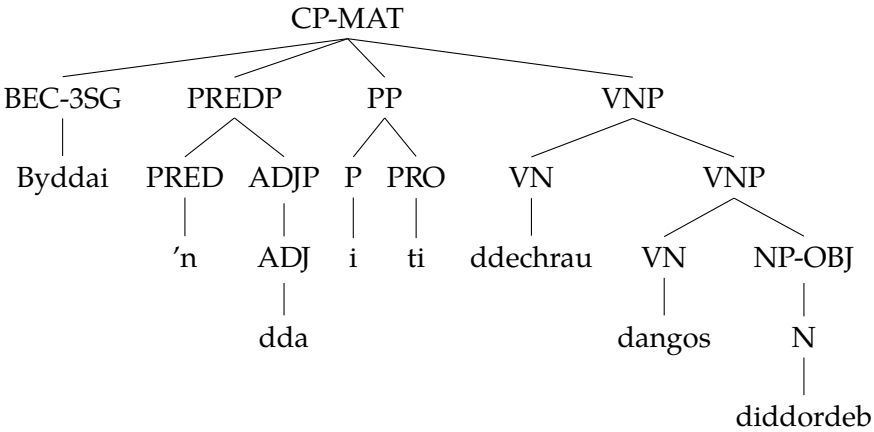
Nonfinite verb phrases in Welsh are built around elements traditionally termed 'verbal nouns' that exhibit infinitive-like behaviour with mixed nominal and verbal properties (Borsley et al. 2007: 68–75). This section focuses on verbal nouns themselves, as well as frequently used constructions where they are preceded by an aspectual marker, for instance, in verbal periphrases involving the progressive particle *yn* or the perfective particle *wedi*. In the previous section, it was concluded that an additional VP layer should be avoided for finite verb phrases. The question is whether multiple layers such as Aspectual and/or Verb Phrases with nonfinite verbs would be appropriate and helpful for users of the PARSHCWL treebank.

In addition to certain general benefits of flatter structures (see introduction to section 2), another reason to avoid additional VP layers for nonfinite verbs would be to create greater parallelism in corpus annotation: if finite verbs are not contained within VPs, then neither should nonfinite verbs be. If layers of aspectual phrases (ASPPs) or VPs were to be created for nonfinite but not for finite verbs, that would result in more complex queries. When looking for direct objects, for example, the query would need to involve searches on (at least) two different levels: within `CP-MAT` and within `ASPP/VP`.

However, multi-level queries are no longer an issue with the more advanced querying tools that are currently available (e.g. later versions of CorpusSearch) or custom-made scripts that can efficiently find objects on any level. Furthermore, to avoid a potential lack of parallelism, instead of a `VP` label in specific nonfinite verbal contexts only, another label could be employed. In the historical English corpora, for example, the label `IP-INF` is used for infinitival clauses. PARSHCWL adopts a similar strategy of adding a nonfinite verbal layer, but uses the label Verbal Noun Phrase (`VNP`) to reflect the distinctive nature of Welsh verbal nouns. Thus we have the following parse for a present-day Welsh example with multiple verbnouns:

(10) *Byddai 'n dda i ti ddechrau dangos diddordeb.*
be.VN PRED good to you begin.VN show.VN interest
'It would be good for you to start showing interest.' (Present-day Welsh)

```
[CP-MAT [BEC-3SG Byddai]
        [PREDP [PRED 'n] [ADJP [ADJ dda]]]
        [PP [P i] [PRO ti]]
        [VNP [VN ddechrau]
            [VNP [VN dangos]
                [NP-OBJ [N diddordeb]]]]]
```

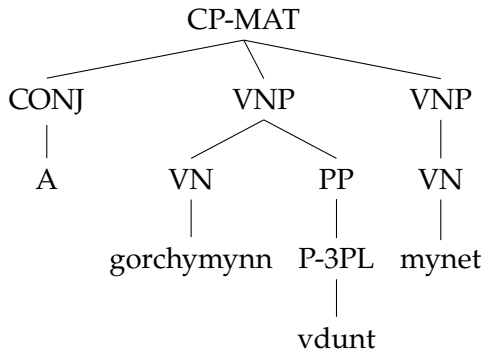


Provided labelling is carried out consistently, querying remains unproblematic. A search for direct objects, for instance, can be limited to elements lying within `CP*` and `VNP` domains.

In Middle Welsh, verbal nouns can also be used as the main verb replacing a finite form in continuous narrative contexts. For the sake of consistency, these also project a `VNP`:

(11) *a gorchymynn vdunt mynet*
and order.VN to.3PL go.VN
'and [he] order[ed] them to go' (RhG, Jesus 119, folio 112v, lines 8–9, Ystoria Adrian ac Ipotis, 14th *c.*)

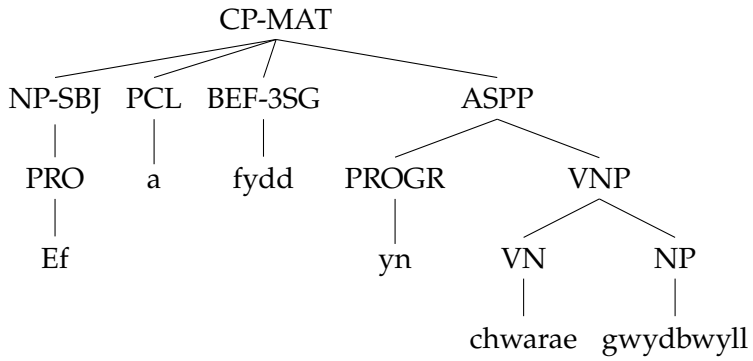
```
[CP-MAT [CONJ A]
        [VNP [VN gorchymynn]
             [PP [P-3PL vdunt]]]
        [VNP [VN mynet]]]
```



Finally, as seen above (section 1.2), aspectual markers preceding nonfinite verbs are taken to head an Aspectual Phrase ASPP. Both nominal and pronominal objects are contained within the VNP. This allows verbs with nominal and pronominal objects to be searched for in the same way: both are instances of ASPP. The constructed Middle Welsh examples (12) and (13) demonstrate.

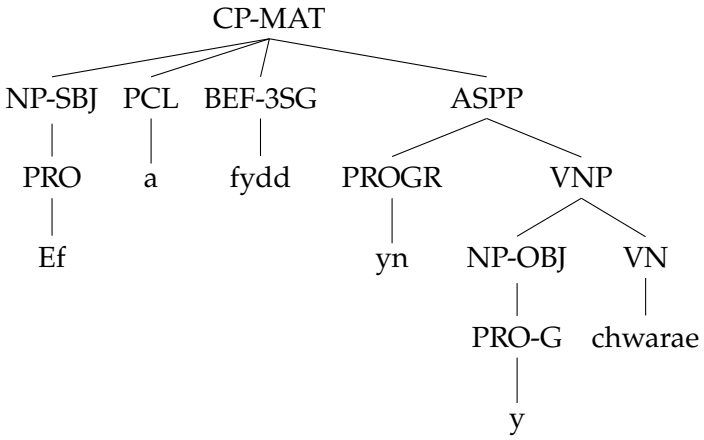
(12) *Ef a fydd yn chwarae gwydbwyll.*
he PRT be.FUT.3SG PROG play.VN gwyddbwyll
'He will be playing gwyddbwyll/chess.'

```
[CP-MAT [NP-SBJ [PRO Ef]]
        [PCL a]
        [BEF-3SG fydd]
        [ASPP   [PROGR yn]
                [VNP    [VN chwarae]
                        [NP gwydbwyll]]]]
```



(13) *Ef a fydd yn y chwarae.*
he PRT be.FUT.3SG PROG 3MSG play.VN
'He will be playing it.'

```
[CP-MAT [NP-SBJ [PRO Ef]]
        [PCL a]
        [BEF-3SG fydd]
        [ASPP   [PROGR yn]
                [VNP [NP-OBJ [PRO-G y]]
                     [VN chwarae]]]]
```



This section has proposed two different strategies for the syntactic annotation of verb phrases in PARSHCWL. For finite verbs, the addition of VP layers was rejected, both because they cannot easily be added automatically, and because they are not particularly helpful for users querying the corpus. A fixed VP layer containing the finite verb (and possibly preverbal particles and pronouns depending on them) would pose various problems for Middle Welsh V2 structures in particular, forcing annotators to undertake too much syntactic analysis that would depend on theory-specific assumptions. For nonfinite verbs, however, these issues do not arise, and it was seen that adding specific layers, labelled `ASPP` and `VNP`, offers more insight into the structure, but does not hinder querying, for instance, for direct objects. Although specific details differ from corpus to corpus (e.g. the `VNP` label for verbal noun phrases is unique to PARSHCWL), the core decisions on when to add and when to avoid additional layers in the verbal domain are broadly in line with those made for the other Penn historical corpora.

3 Complex Prepositional Phrases

Welsh, like many other languages, has complex prepositional phrases (PPs) and adverbs, many of which grammaticalized from sequences of preposition + noun, for example, *ar ben*, 'on top of' < *ar* 'on' + *pen* 'head'. The treatment

of multi-word prepositions like these has been an issue for other Penn historical corpora. Thus, for instance, Booth, Breitbarth, Ecay & Farasyn (2020) treat the individual elements of such structures in Middle Low German either as prepositions or nouns according to a conservative grammatical analysis, leaving further research to the user. In other contexts, namely the categorization of entire clauses, they introduce a convention for labelling indeterminate items as `X`, a solution similar in spirit to the one we shall advocate for Welsh prepositional phrases here.

Welsh can combine not only two, but also three originally separate items. One of these complex constructions also has inflected forms with a pronominal object, namely *ar hyd* 'along (on [the] length [of])', hence *ar hyd-ddo* 'along it'. These structures raise issues about what to do with an element that has more than one plausible parse, especially when the two possibilities may both be justified for different stages of the history of the language. These are both topics attracting a good deal of interest in recent literature on the construction of parsed historical corpora; on issues of annotational uncertainty inherent in linguistic structure, see Barteld, Ihden, Schröder & Zinsmeister (2014) and Beck, Booth, El-Assady & Butt (2020), and on the issues posed by change, albeit within a framework of Construction Grammar and fuzzy categories, see Merten & Seemann (2018). This section deals with the question of how to provide consistent morphosyntactic annotation for the most complex prepositions and adverbs in the corpus.

3.1 Preprocessing and POS tagging basic PP elements

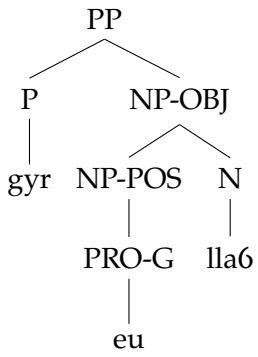
Treatment of complex prepositions raises questions that span preprocessing (regularization of word boundaries), POS tagging and syntactic description. In PARSHCWL, the POS tags used for complex prepositions reflect their combined nature: the combined preposition *ar ben* 'on top of', for example, is tagged `ar/P ben/PX`, where `PX` both reflects the fact that *ben* is part of a combined preposition, and that it is the second part, which is not clearly itself a preposition (in this case it is an etymological noun that may or may not have completed the transition to become part of a preposition). This section sets out how this decision was reached, building on the preprocessing protocols discussed in section 1.1 above.

Consider a typical case such as (14).

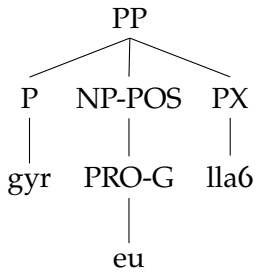
(14) *gyr* *eu* *lla6*
by 3PL hand
'near them' (RhG, Peniarth 4, folio 17v, 68.30–31, 14th *c.*, *Manawydan*)

Two plausible structural descriptions of this phrase are given in (15) and (16). In (15), it is treated as a prepositional phrase with a noun phrase complement, and *llaw* is labelled a noun, while (16) offers a flatter structure with no internal hierarchy, and the category of *llaw* is left open by using a dedicated tag limited to this construction, namely `PX`.

(15) `[PP [P gyr]`
`[NP-OBJ [NP-POS [PRO-G eu]]`
`[N lla6]]]`



(16) `[PP [P gyr]`
`[NP-POS [PRO-G eu]]`
`[PX lla6]]`



PARSHCWL opts for the second of these. The motivation behind this decision is that it allows all potential cases of complex prepositions to be extracted easily by users of the corpus, and that it avoids the need to express an opinion on the status of the etymological noun (or similar element) for any given example. Although there is a substantial set of complex prepositions and adverbs in Welsh today, individual items have grammaticalized gradually and at different times. In some cases, grammaticalization is fully complete today and an item's origin may no longer be evident. This may be either because the element at its core has become obsolete in other functions (e.g. *ar gyfer* 'for' or *o herwydd* 'because of', where *cyfair*/*cyfer* 'direction' and *herwydd* 'cause, reason' are largely obsolete as nouns), or because there has been semantic divergence (e.g. *ar bwys* 'next to' is no longer obviously related to *pwys* 'weight'). Conversely, at earlier periods, the linguist may conclude that grammaticalization has not yet begun, and a phrase that later develops into a complex preposition should be analysed solely in terms of its component parts. It is difficult to pinpoint exactly when or even whether such elements

stopped being nouns and started being part of a complex preposition, and, indeed, this could be a fruitful topic of research based on the treebank. In order to avoid taking a stance on this question for individual cases, the tag PX is created for second and subsequent elements of potential complex prepositional phrases. This allows users to extract all relevant instances easily. It is then up to them to decide how to deal with them in terms of their own analysis.

This approach entails defining an exhaustive list of potential complex prepositions. This was begun using those combinations listed by Evans (1964: 181–220). However, because this is limited to medieval Welsh, and the corpus aims for historical continuity into the modern period, additions based on items listed in the University of Wales Dictionary (*Geiriadur Prifysgol Cymru*, *GPC*) had to be considered; for discussion, see below. In cases of doubt, PX is preferred. A full list will be included in the POS-tagging section of the PARSHCWL annotation manual.

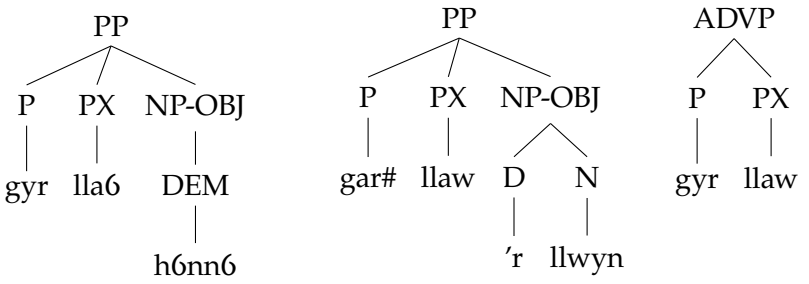
In the examples considered so far, the two parts of the complex preposition are separated by a pronominal object and thus have to be tagged as two separate elements. Should we adopt the same approach in cases where they are adjacent, that is, when a potential complex preposition is used with a lexical object, for instance, Modern Welsh *gerllaw'r ffordd* 'next to the road', or is used adverbially with no object at all, as with *gerllaw* 'nearby, at hand'? In these cases, orthography (whether the item is spelled as one or two words) might be thought to offer a guide. However, in practice, it is of little use, and the wide range of orthographic practice attested in pre-modern Welsh is more of a hindrance than a guide to data collection and analysis. Potential complex prepositions may historically be written as two separate words or as one, thus, in our current example as either *gyr llaw* or *gyrllaw*. In Modern Welsh, some items, as is the case for *gerllaw*, are conventionally written as one word, while others are conventionally written as two. If written as one word, *gerllaw* could simply be tagged as a preposition P or adverb ADV. However, since an additional object pronoun, as in (14) above, would break up this jointly written complex preposition, it is more consistent to take a uniform approach and to split such combinations and tag each element separately even when they are written as one word. Jointly written *gerllaw* is thus split (the split marked with #) during the preprocessing stage into *ger llaw*, and is tagged ger#/P llaw/PX; see Meelen & Willis (2021) for further details. A selection of such examples with their POS tags and structural descriptions follows:

(17) a. *gyr lla6 h6nn6*
nearby DEM.MSG
'(near)by that one' (RhG, Peniarth 4, folio 8r, 30.22, 14th *c.*)

b. *cyd-ddigwyddaw garllaw 'r llwyn*
together-fall.VN by the grove
'falling down together by the grove' (GDG 202, 14th *c.*)

c. *a chat ym prydein gyr llaw*
and battle in Britain at hand
'and a battle in Britain nearby' (CAP3, p. 117, *c.* 1300)

```
[PP [P gyr]            [PP [P gar#]          [ADVP [P gyr]
    [PX lla6]              [PX llaw]                [PX llaw]]
    [NP-OBJ [DEM h6nn6]]] [NP-OBJ [D 'r]
                                  [N llwyn]]]
```



The PX convention diverges from the practice of some of the other parsed historical corpora. Most other Penn-style corpora treat these elements etymologically. Thus, the Parsed Corpus of Early English Correspondence (PCEEC) and the Penn Parsed Corpus of Modern British English (PPCMBE2) tag *in his stead* as `in/P his/PRO$ stead/N` and *in stead of him* as `in/P stead/N of/P him/PRO`, and use a composite tag for *instead* written as one word as `instead/P+N`. PARSHCWL avoids composite tags due to their potential complexity for Middle Welsh and their potential to disrupt easy searching. The current approach also avoids the danger of wildly anachronistic tags that could arise if an etymological tag was maintained long after a grammaticalization process had been completed. For instance, it could have forced use of `N` as the tag for *gyfer* in *ar gyfer* 'for' for Modern Welsh, an etymologizing tag that native speakers would not find at all intuitive today.

Another possibility would have been to follow the Reference Corpus of Middle Low German / Low Rhenish (1200–1650) (Schröder 2014, Peters & Nagel 2014), which indicates the derivation or inflection of all items at different levels of analysis; thus, *bekant* 'known' is `ADJN<VVPP` ('nominative adjective' from 'past participle of a verb' and *blîven* 'remained (3sg. past)' is `VVFIN<VV` ('finite verb' from 'verb'). This would open up the possibility of using `PX<N`, a prepositional extension derived from a noun. This, however, would be time-consuming to implement across all words of the corpus, and,

in any case, would still mean tracing the etymology of every element, potentially leading to multiple tags and/or disagreement over contentious forms.

A good example in this regard is Middle Welsh *y rwg* 'between'. Etymologically, this is generally agreed to be from *er* 'near, by' (cognate with Latin *per*) plus *wng*/*wnc* 'nearness, proximity' (Morris Jones 1913: 405; Sims-Williams 2013: 33). Thus, an etymological tag would be `y!r/P wg/N`, involving reassigning the word boundary (by convention marked using `!`) and assigning the tag `N` to *wg*, an element not found independently in any historical variety of Welsh. By using the current convention, we can avoid bringing such considerations into the corpus, and simply assign a fairly neutral sequence of tags, namely `y/P rwg/PX`.

`PX` is thus relatively straightforward to implement, and is in line with the general principles of PARSHCWL, potentially contentious linguistic analyses being avoided with a generic label.

3.2 Tagging problematic cases of complex PPs

Some potential cases of complex prepositions in Welsh are problematic because they do not appear among those given by Evans (1964), who is concerned solely with the medieval period. These furthermore sometimes manifest a number of variants. For instance, the noun *gwrthwyneb* 'opposition' appears in several variant phrases. Compare example (18) with *yn* 'in' (fused with the noun in *yngwrthwyneb*) and *i* 'to', with (19), where only *yn* is present.

(18) *yngwrthwyneb* *i* *hynn*
in.opposition to DEM.PROX.NSG
'contrary to this' (HCWL, *Perl mewn adfyd* 58.19, 1595)

(19) *yngwrthwyneb* *hynny*
contrary DEM.DIST.NSG
'contrary to this' (HCWL, Peniarth 218 *Mandefil*, line 731, 1610)

While *yngwrthwyneb* is written as one word in these examples, it is also found as two words, the standard spelling today:

(20) *ac* *edrych* *yn* *g6rth6yneb* *y* *t6r6f* *a* *orugant*
and look.VN in contrary the commotion PRT do.PST.3PL
'And they looked in the direction of the commotion.' (RhG, Peniarth 4, folio 64r, 389.41–42, 14th *c.*, *Gereint*)

Finally it should be noted that, while this collocation is found as early as Middle Welsh, it is not listed by Evans (1964: 216–19), although it is listed in the University of Wales Dictionary.

A similar issue arises with the preposition that in Present-Day Welsh is *gyda(g)* '(together) with'. In addition to spelling variants, this occurs historically also as *y gyd a*, which is the main form in Middle Welsh. The noun that grammaticalized as part of this is *cyd* 'union, combination' and the following element is the preposition *a(c)* 'with'. The origin of the first element *y* is not entirely certain, but it is fairly likely to be the preposition *i* (Middle Welsh *y*) 'to', found also in the adverb *i gyd* 'all (together)'. The two main historical forms are given in (21) and (22).

(21) *y gyt* *ac* *wynt*
together with 3PL
'together with them' (PKM 4.14, 14th *c.*)

(22) *gyd* *a* *chwi*
together with 2PL
'(together) with you' (HCWL, *Testament Newydd*, Matt. 26:11, 1567)

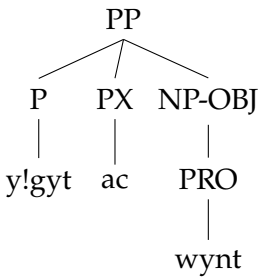
To make parsing more straightforward, during the preprocessing stage, both of these examples, and cases like them, could be combined to form a single preposition, using the ! convention for joining words written separately, that is, as `yngwrthwyneb!i/P` 'contrary to' and `y!gyd!a/P` 'with'. However, in the former case, there is the possibility of a genitive pronoun breaking up this unit. Also, because of the range of orthographic variants, it would also not be particularly straightforward to do this.

In the case of *yng ngwrthwyneb*, there is also the question of whether the item has conventionalized sufficiently to be considered a complex preposition. Clearly, the answer to this question varies according to the historical period under consideration, and the existence of variants would be evidence for a negative answer to this question.

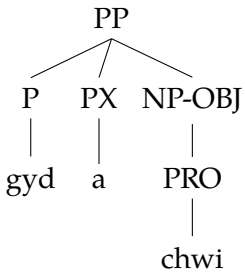
We therefore rejected this approach that combines them into a single preposition. Ultimately, it was considered preferable to use the `PX` tag, adopting a generous approach to its application, so as to allow anyone searching for complex prepositions to extract all possible instances. In this way, users can decide for themselves the degree of grammaticalization at any given point in time. This necessitates an extension to the original list of items to which the `PX` convention is applied.

A structure with `PX` tags like the one presented in (23) is therefore preferred in all these cases:

(23) `[PP [P y!gyt]`
`[PX ac]`
`[NP-OBJ [PRO wynt]]]`



(24) `[PP [P gyd]`
`[PX a]`
`[NP-OBJ [PRO chwi]]]`

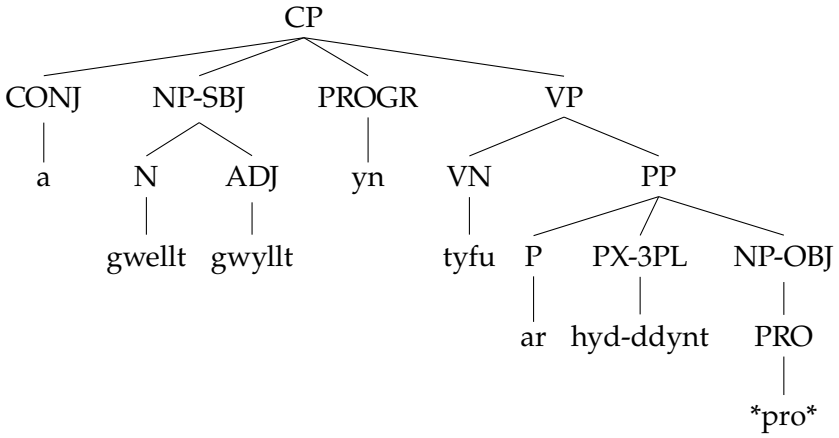


A final complication arises with the inflecting complex preposition *ar hyd* 'along':

(25) *a* *gwellt* *gwyllt* *yn* *tyfu* *ar* *hyd-ddynt*
and grass wild PROG grow.VN on length-3PL
'with wild grass growing along them' (HCWL, *Yr American* 6.34, 1840)

This can be dealt with by allowing the `PX` element to have a person–number feature added to it, something which would have been less intuitive if the element had retained its etymological parse as a noun `N`. As with simple inflected prepositions, structures like these need to include null pronouns that are added in the parsed treebank as `*pro*`. Example (25) is thus parsed as in (26). As in the Tycho Brahe Parsed Corpus of Historical Portuguese (TBCHP), the function of a null pronoun is marked by putting it inside a noun phrase with a defined grammatical function.

(26)
```
[CP [CONJ a]
    [NP-SBJ [N gwellt]
            [ADJ gwyllt]]
    [PROGR yn]
    [VP [VN tyfu]
        [PP [P ar]
        [PX-3PL hyd-ddynt]
        [NP-OBJ [PRO *pro*]]]]]
```



4 Coordination

We now turn to the third case that we will consider, namely clausal coordination.

As we have seen, Middle Welsh is a V2 language, in which a canonical main clause begins with a phrase (e.g. subject, object, adverbial, nonfinite verb etc.) followed by a particle indicating the grammatical function of this initial phrase and then the finite verb. The initial phrase has generally been argued in the literature to amount to a topic in information-structure terms (Fife 1988, Poppe 1991, Willis 1998, Meelen 2016), and the structure is thus a way to implement a form of topic–comment word order. We use the term 'topic' to refer to this initial element, accepting of course that its precise nature is the potential subject of further research using the corpus.

This word-order system produces some special issues when elements are omitted from the second conjunct under some loose form of identity with elements in the first conjunct. The possibilities attested in historical Welsh go beyond those found in historical English and some of the other languages for which historical Penn-based corpora have been created, motivating certain additions to the parsing conventions. For fuller details of the linguistic con-

straints, see Willis (1997, 1998: 103–122) and Borsley et al. (2007: 299–302).

4.1 Null topics

Where two full clauses are conjoined, as in (27), no particular issues are encountered: each clause is represented as a full clause according to the usual conventions, and the second of them is introduced by *a*(*c*) 'and', tagged as CONJ.

(27) *mi a wnn pwy wytti, ac ny chyuarchaf*
I PRT know.PRS.1SG who be.PRS.2SG you and NEG greet.PRS.1SG
i well it.
I better to.2SG
'I know who you are, and I shall not greet you.' (PKM 2.8, 14th *c.*)

However, problems arise when the second clause is introduced by a particle and an element that is shared between the two clauses is omitted in the second clause. For instance, in the following example, the subject *mi* 'I' is shared by both clauses, but is not directly represented in the second clause; its presence can only be inferred from the clause-initial particle *a*, which indicates that the preceding element is a subject or an object.

(28) *mi a rodaf Pryderi a Riannon it*
I PRT give.PRS.1SG Pryderi and Rhiannon to.you
'I shall give Pryderi and Rhiannon to you'

ac __ a waredaf yr hut a 'r lletrith y
and TOP PRT remove.PRS.1SG the spell and the magic from
ar Dyuet.
on Dyfed.
'and remove the spell and magic from Dyfed.' (PKM 64.19–20, 14th *c.*)

In cases such as these, should we represent the omitted element at the start of the second clause in some way? The argument for doing so is that it can be any of a wide range of elements and it is therefore not easy to search for particular types without additional assistance; in fact, the omitted element may be anything that can appear before a preverbal particle in Middle Welsh. Other examples are given below, showing the range of possibilities. In example (29), the object *a gauas o achenogyon* [...] 'such needy people as he could find [...]' is fronted in the first conjunct, and this is understood as the object in the second conjunct; the preverbal particle *a* indicates that the element

that precedes it is a subject or an object; hence, we understand there to be an omitted object topic immediately before it.

(29) *a* *gauas* *o* *achenogyon* *yn* *y* *holl* *lu* *a*
REL find.PST.3SG of needy.PL in his whole host PRT
wisgwys *yn* *hard*
clothe.PST.3SG PRED beautiful
'such needy people as he could find in his whole host he clothed beautifully'

ac __ *a* *borthes* *yn* *enrydedus*
and TOP PRT feed.PST.3SG PRED honourable
'and fed honourably.' (YCM 21.32–22.2, 14th *c.*)

In example (30), the second conjunct begins with the preverbal particle *y*(*d*), which normally co-occurs with a preceding adverbial topic. We can therefore posit that an adverbial topic (semantically, either 'every day at midday' or a semantically bleached 'then') is understood at the start of the second conjunct.

(30) *peunyd* *pob* *hanher* *dyd* *y* *kymerei* *y* *deu*
every.day every mid day PRT take.IMPF.3SG the two
amherawdyr *eu* *bwyt*
emperor their food
'every day at midday the two emperors would have their food'

ac __ *y* *peidynt* *ac* *ymlad*
and TOP PRT stop.IMPF.3PL with fight.VN
'and stop fighting.' (BMW lines 278–279, 14th *c.*)

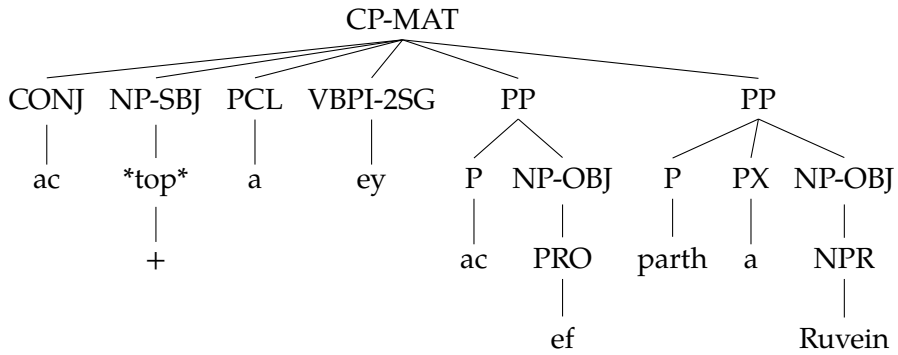
Finally, example (31) shows that the omitted element in the second clause need not have been in the topic position in the first clause. There, the omitted element is a second person singular pronoun 'you' (as indicated by agreement on the verb *ey* 'go'), which was a postverbal subject in the first conjunct clause. In fact, the omitted topic may have played virtually any role in the first conjunct and may have been in preverbal topic position or in a postverbal position.

(31) *A* *'r* *mab* *hwnnw* [...] *a* *aberthy* *di* *y* *Duw,*
and the boy DEM.MSG PRT sacrifice.PRS.2SG you to God
ac __ *a* *ey* *ac* *ef* *parth* *a* *Ruvein.*
and TOP PRT go.PRS.2SG with him towards with Rome

> 'And that boy [...] you will sacrifice to God and [you] will go with him towards Rome.' (KAA lines 20–22, 14th *c.*)

A consistent treatment across all such cases is desirable to ease searching. Consequently, a topic element is inserted in all cases, given as + (the convention for 'missing character'; see section 1.1 above) in the text, tagged for part of speech as `*top*` (topic), and contained inside a phrase that identifies its category and grammatical function, thus `NP-SBJ` in (28) and (31), `NP-OBJ` in (29), and `AP` in (30). This approach differs somewhat from that adopted in the Penn English historical corpora, where an element `*con*` is used to indicate a subject elided in coordination (e.g. *He arrived and* `*con*` *sat down*). This difference in approach is motivated by the much wider range of possibilities attested in Middle Welsh as compared to Middle English, and by the extensive interaction with the verb-second system of word order. While example (28) could be handled using the `*con*` convention, some additional machinery is needed to cope with the other examples quoted. The structural description for the second clause of (31) is therefore:

(32)
```
[CP-MAT [CONJ ac]
        [NP-SBJ [*top* +]]
        [PCL a]
        [VBPI-2SG ey]
        [PP [P ac] [NP-OBJ [PRO ef]]]
        [PP [P parth] [PX a] [NP-OBJ [NPR Ruvein]]]]
```



As in the Penn corpora, the initial coordination marker *a*(*c*) 'and' is tagged `CONJ` and treated as part of the second conjunct clause.

A further issue arises with agreement in certain cases. Agreement patterns show that the omitted topic of the second conjunct need not always correspond to a single grammatical element in the first conjunct. This is the case in (33), where the verb in the second conjunct is plural, and its subject must

therefore be interpreted not as continuing the subject topic of the first conjunct, namely *hwnnw* 'that one, he [*sc.* the giant Llasar Llaes Gyfnewid]', but as him and his wife. While cases such as these could be searched for by checking for verb-agreement patterns, it nevertheless seems useful to mark number on omitted topics to facilitate this. Thus, in the second conjunct of (33), the omitted topic receives the tag `*top-pl*`.

(33) *A hwnnw a doeth yma o Iwerdon, a*
and DEM.MSG PRT come.PST.3SG here from Ireland and
Chymidei Kymeinuoll, y wreic, y gyt ac ef
Cymidei Cymeinfoll his wife together with him
'And he came here from Ireland, with Cymidei Cymeinfoll, his wife, with him'

ac __ a dianghyssant o 'r ty hayarn yn Iwerdon
and TOP PRT escape.PST.3PL from the house iron in Ireland
'and [they] escaped from the iron house in Ireland.' (PKM 35.5–7, 14th *c.*)

Finally, an expletive subject, third-person singular masculine pronoun *ef* 'he, it', may appear in the topic position of such clauses (cf. similar uses in other V2-languages, as with German *es*, Dutch *er*, Icelandic *það*, Swedish *det* etc.; see Vikner 1995). There is the question of what to do with expletive pronouns in these sentences. Expletive subjects can be shared across two conjoined clauses, with the expletive subject present in the topic position of the first conjunct and omitted in the second:

(34) *ef a vuwyt lawen wrthaw,*
3MSG PRT be.PST.IMPERS happy at.3MSG
'people were happy towards him' (lit. 'it/there was happy')

ac __ a gymerwyt y uarch o 'e ystablu
and EXP PRT take.PST.IMPERS 3MSG horse to 3MSG.GEN stable.VN
'and his horse was taken to be stabled' (lit. 'and it/there was taken his horse to be stabled') (YSG lines 1770–71, 14th *c.*)

As before, the two clauses do not have to be parallel, and the gap in the second clause can represent an expletive subject even if there was no overt expletive subject in the first conjunct. Thus, in (35), the second and third conjuncts have a gap before the particle. The particle is *a*, which indicates a preceding subject or object. Impersonal clauses such as these may elsewhere

be introduced by expletive *ef* (cf. the first conjunct of (34) above). Since these clauses are otherwise complete (the verb is impersonal, and the direct object in both cases is *'n* 'us'), and the particle *a* indicates a preceding subject or object, we must conclude that there is an unexpressed expletive subject before the particle in each clause.[5]

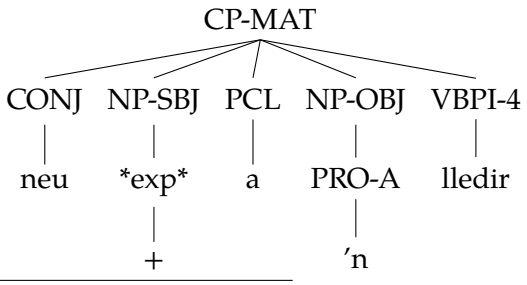
(35) *ni* *a* *gollwn* *ynn* *kyuoeth*
we PRT lose.PRS.1PL our wealth
'We will lose our wealth'

ac __ *a* *'n* *carcherir* *yny* *vom*
and EXP PRT 1PL.ACC incarcerate.PRS.IMPERS until be.PRS.SBJV.1PL
ueirw
dead.PL
'and [we] will be incarcerated until we die'

neu __ *a* *'n* *lledir.*
or EXP PRT 1PL.ACC kill.PRS.IMPERS
'or [we] are killed.' (14th *c.*, BTy-RBH 66.13–14)

Given this logic, and given the desirability of maintaining parallel treatment across all instances of coordination, we mark a gap here and indicate its function, the tag being `*exp*` and the phrase marker being `NP-SBJ`:

(36)
```
[CP-MAT [CONJ neu]
        [NP-SBJ [*exp* +]]
        [PCL a]
        [NP-OBJ [PRO-A 'n]]
        [VBPI-4 lledir]]
```



5 The Icelandic IcePaHC corpus inserts an element `*exp*` in older Icelandic wherever an overt expletive would be required in the language today. Use of the `*exp*` convention is very much in line with what we are proposing for Welsh, although the specific approach of comparing earlier and later stages of the language cannot be directly transposed, since Welsh lost expletive subjects at the same time that it lost its V2-system; see Willis (1998: 181–256).

5 Conclusion

The case studies presented in this article have illustrated the multiplicity of issues that need to be decided on in the construction of any parsed corpus. The decisions that were ultimately adopted were guided by a number of core principles: the need for the corpus to be of maximal use as a searching device for users; the consequent desire to make the structures adopted theory-neutral to the extent that this is possible; the need to adopt conventions that are robust across large historical time periods even for a changing language; and the need for conventions to be exhaustive enough that both annotators and users can be confident of the annotation to be adopted in all instances. While these decisions can never be perfect, their adoption represents an important step on the way to constructing the final corpus.

List of corpora and other primary sources

BMW	*Breudwyt Maxen Wledic*, ed. Brynley F. Roberts. Dublin: Dublin Institute for Advanced Studies, 2005.
BTy-RBH	*Brut y Tywysogyon or The Chronicle of the Princes: Red Book of Hergest Version*, ed. Thomas Jones. Cardiff: University of Wales Press, 1955.
CAP3	Y Cyfoesi a'r Afallennau yn Peniarth 3, ed. Ifor Williams. *Bulletin of the Board of Celtic Studies* 4.112–129, 1927.
GDG	*Gwaith Dafydd ap Gwilym*, ed. Thomas Parry. Cardiff: Gwasg Prifsygol Cymru, 1952.
HCWL	A Historical Corpus of the Welsh Language 1500–1850, ed. David Willis & Ingo Mittendorf, www.celticstudies.net, accessed 23 October 2021.
IcePaHC	Icelandic Parsed Historical Corpus, ed. Joel C. Wallenberg, Ingason, Anton Karl, Einar Freyr Sigurðsson & Eiríkur Rögnvaldsson, www.linguist.is/icelandic_treebank, accessed 23 October 2021.
KAA	*Kedymdeithyas Amlyn ac Amic*, ed. Patricia Williams. Cardiff: Gwasg Prifysgol Cymru, 1982.
MCH	James, C. 2013. *Machlud Cyfraith Hywel: Golygiad o BL Add. 22356*, ed. Christine James. Cambridge: Seminar Cyfraith Hywel, 2013. https://cronfa.swan.ac.uk/Record/cronfa14807, accessed 23 October 2021.
PCEEC	Parsed Corpus of Early English Correspondence, ed. Arja Nurmi & Ann Taylor, https://varieng.helsinki.fi/CoRD/corpora/CEEC/pceec.html, accessed 23 October 2021.

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

PPCMBE2 Penn Parsed Corpus of Modern British English, ed. Anthony Kroch, Beatrice Santorini & Ariel Diertani, www.ling.upenn.edu/ppche/ppche-release-2016/PPCMBE2-RELEASE-1, accessed 23 October 2021.

PPCME2 The Penn–Helsinki Parsed Corpus of Middle English, ed. Anthony Kroch & Ann Taylor, www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-4/index.html, accessed 23 October 2021.

RhG Rhyddiaith Gymraeg 1300–1425, ed. Diana Luft, Peter Wynn Thomas & D. Mark Smith, www.rhyddiaithganoloesol.caerdydd.ac.uk, accessed 23 October 2021.

TBCHP Tycho Brahe Parsed Corpus of Historical Portuguese, ed. Charlotte Galves, Aroldo Leal de Andrade & Pablo Faria, www.tycho.iel.unicamp.br/~tycho/corpus/texts/psd.zip, accessed 23 October 2021.

YCOE The York–Toronto–Helsinki Parsed Corpus of Old English Prose, ed. Ann Taylor, Anthony Warner, Susan Pintzuk & Frank Beths, https://www-users.york.ac.uk/~lang22/YcoeHome1.htm, accessed 23 October 2021.

YCM *Ystorya de Carolo Magno*, ed. Stephen J. Williams. Cardiff: Gwasg Prifysgol Cymru, 1930.

YSG *Ystoryaeu Seint Greal*, ed. Thomas Jones. Cardiff: Gwasg Prifysgol Cymru, 1992.

References

Barteld, Fabian, Sarah Ihden, Ingrid Schröder & Heike Zinsmeister. 2014. Annotating descriptively incomplete language phenomena. In Lori Levin & Manfred Stede (eds.), *Proceedings of LAW VIII: The 8th Linguistic Annotation Workshop*, 99–104. Dublin: Association for Computational Linguistics and Dublin City University.

Beck, Christin, Hannah Booth, Mennatallah El-Assady & Miriam Butt. 2020. Representation problems in linguistic annotations: Ambiguity, variation, uncertainty, error and bias. In Stephanie Dipper, Amir Zeldes, Luke Gessler & Adam Roussel (eds.), *Proceedings of the 14th Linguistic Annotation Workshop*, 60–73. Barcelona: Association for Computational Linguistics.

Booth, Hannah, Anne Breitbarth, Aaron Ecay & Melissa Farasyn. 2020. A

Penn-style treebank of Middle Low German. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 766–775. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.96.

Borsley, Robert, Maggie Tallerman & David Willis. 2007. *The syntax of Welsh*. Cambridge: Cambridge University Press.

Chernodub, Artem, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann & Alexander Panchenko. 2019. TARGER: Neural argument mining at your fingertips. In *Proceedings of the 57th Annual Meeting of the Association of Computational Linguistics (ACL'2019)*, Florence, Italy.

Daelemans, Walter, Jakub Zavrel, Antal Van den Bosch & Ko Van der Sloot. 2010. Mbt: Memory-based tagger - version 3.2 Reference Guide. http://ilk.uvt.nl/downloads/pub/papers/ilk.1004.pdf.

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

Fife, James. 1988. *Functional syntax: A case study in Middle Welsh*. Lublin: Redakcja Wydawnictw Katolickiego Uniwersytetu Lubelskiego.

Meelen, Marieke. 2016. *Why Jesus and Job spoke bad Welsh: The origin and distribution of V2 orders in Middle Welsh*. Utrecht: LOT Dissertation Series.

Meelen, Marieke & David Willis. 2021. Towards a historical treebank of Middle and Early Modern Welsh, part I: Workflow and POS tagging. *Journal of Celtic Linguistics* 10. 125–154.

Merten, Marie-Luis & Nina Seemann. 2018. Analyzing constructional change: Linguistic annotation and sources of uncertainty. In Francisco José García-Peñalvo (ed.), *TEEM'18: Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*, 819–825. New York/Salamanca: Association for Computing Machinery. doi:doi.org/10.1145/3284179.3284320.

Morris Jones, John. 1913. *A Welsh grammar: Historical and comparative*. Oxford: Clarendon Press.

Peters, Robert & Norbert Nagel. 2014. Das digitale 'Referenzkorpus Mittelniederdeutsch / Niederrheinisch (ReN)'. *Jahrbuch für Germanistische Sprachgeschichte* 5. 165–175. doi:doi.10.1515/jbgsg-2014-0012.

Poppe, Erich. 1991. *Untersuchungen zur Wortstellung im Mittelkymrischen*. Hamburg: Helmut Buske Verlag.

Randall, Beth, Ann Taylor & Anthony Kroch. 2005. Corpussearch 2. http://corpussearch.sourceforge.net/.

Santorini, Beatrice. 2022. Annotation manual for the Penn Parsed Corpora of Historical English and the Parsed Corpus of Early English

Correspondence. https://www.ling.upenn.edu/~beatrice/corpus-ling/annotation-202x/.

Schröder, Ingrid. 2014. Das Referenzkorpus: Neue Perspektiven für die mittelniederdeutsche Grammatikographie. *Jahrbuch für Germanistische Sprachgeschichte* 5. 150–164. doi:doi.10.1515/jbgsg-2014-0011.

Sims-Williams, Patrick. 2013. Variation in Middle Welsh conjugated prepositions: Chronology, register and dialect. *Transactions of the Philological Society* 111(1). 1–50.

Vikner, Sten. 1995. *Verb movement and expletive subjects in the Germanic languages*. Oxford: Oxford University Press.

Willis, David. 1997. Clausal coordination and the loss of verb-second in Welsh. *Oxford Working Papers in Linguistics, Philology and Phonetics* 2. 151–172.

Willis, David. 1998. *Syntactic change in Welsh: A study of the loss of verb-second*. Oxford: Oxford University Press.

Willis, David & Ingo Mittendorf. 2004. Ein historisches Korpus der kymrischen Sprache. In Erich Poppe (ed.), *Keltologie heute: Themen und Fragestellungen*, 135–142. Münster: Nodus.

Marieke Meelen
University of Cambridge
Trinity Hall
Trinity Lane
Cambridge
CB2 1TJ
mm986@cam.ac.uk

David Willis
University of Oxford
Jesus College
Turl Street
Oxford
OX1 3DW
david.willis@ling-phil.ox.ac.uk
davidwillis.net