Challenges, opportunities and insights annotating historical Welsh

Marieke Meelen

Leiden University Centre for Linguistics

1 Introduction

With five tenses (Future, Past, Pluperfect, Present Subjunctive and Conditional/Imperfect) and 7 different person-number suffixes each, Literary Welsh (as found in most historical texts from the 9th century onwards) has a “rich Romance-like” tense inflection (cf. Roberts (2010)). Furthermore, in Welsh, just like in Irish or Breton, prepositions can be inflected for person, number and gender. The first written evidence of the Welsh language (some glosses and marginal poems) dates from the Old Welsh period (c. 800-1150). The Middle Welsh period, 1150-1500, is more interesting from a syntactic point of view, because for the first time, extensive Welsh tales and chronicles were written down. In the 16th century, the word order of the Welsh language changed from a preferred V2 to the current Modern Welsh VSO order (cf. amongst others Borsley, Tallerman, and Willis (2007), Willis (1998)).

In this paper I will discuss some of the challenges I faced annotating this morphologically rich historical corpus, some specific to Welsh (or other Celtic languages), others inherent to working with historical corpora in general. Apart from challenges, there are also promising opportunities for future research. I will argue for a standardised extension of the tagsets currently used for the Historical English and Icelandic corpora, for annotating languages like Welsh with very rich inflection in the form of verbal or prepositional suffixes. Finally, the way to make this available and useful for all researchers is by storing it in XML format, so that it can easily be queried and furthermore extended with any features in the future.

2 Systematically extending the tagset

Since the standard UPenn annotation scheme (cf. www.ling.upenn.edu/histcorpora/annotation) did not provide enough information to answer certain research questions (mainly queries concerning agreement patterns and change in Information Structure), I have extended the Parts-of-Speech tagset. Starting from the already extended tagset used for the Icelandic corpus (cf. Joel C. Wallenberg and Rögvaldsson (2011)), I have compared the features of Middle and Early Modern Welsh grammar and systematically added dash-tag features, mainly in the verbal domain.

2.1 Adding person and number in verbal inflection

Pilot studies of the short (n = 20k tokens) 16th century Cronicl Hywel ap Syr Mathew demonstrate that adding person and number features for verbal suffixes does not yield a significantly lower Global Accuracy using the Memory-Based Tagger (MBT) by Timbl (cf. Daelemans and Van den Bosch (2005)). This tagger showed robust results and flexibility with the highly variable orthography of minimally preprocessed Welsh texts (cf. www.ilk.uvt.nl/mbt for a detailed manual). The parameter settings of MBT allow for focus on the context and the last 3 letters of unknown words. Since Literary 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-ais “I saw” as VBD-1SG denoting ’preterite-1sg’). 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 & Morgan, 1980, 103,107).

2.2 Further extensions of the tagset

Further extensions of the tagset included ADJQ for equative constructions, inflected prepositions like iddi ’to her’ as P-3SGF, a range of particles according to their function (e.g. PCL-QU, PCL-FOC, PCL-NEG), progressive (PROGR) or predicative (PRED) yn, more refined pronoun tags for conjunctive (PROC) and reduplicated (PROR) forms, VN for verbal nouns and some specific Welsh lexical items, like the verbal noun bod ‘to be’ (that can also function as a complementiser). Tests with this very extensive tagset on a relatively small (n = 65k tokens) 14th century Middle Welsh corpus (that was not extensively preprocessed yet) show similar results with a Global Accuracy of around 90%. Less frequent categories, in particular combined tags for merged forms that were not separated during tokenisation in the preprocessing stage, exhibit significantly worse results. This then, appears to be the limit of useful extension of the tagset. Expanding the training set can improve those results slightly, but more extensive (and thus time-consuming) preprocessing will in the end provide more research opportunities.
2.3 Challenges and limitations

Preprocessing in itself can be challenging. Especially if there is only a “hands-off” diplomatic text edition of one single manuscript, tokenisation forces decisions on splitting certain merged combinations, like yr ‘to the’ and ae ‘and his’. This works as long as there is a logical boundary (e.g. yr can be split up in y ‘to’ and r ‘the’), but for some fused forms, it poses more difficulties, e.g. y (from y + y) ‘to his, her’. This problem is further complicated by the fact that y in Middle Welsh can have a variety of meanings, ranging from the definite article to the preposition ‘to’ and various pronominal forms. Preprocessing will thus have to be done manually, to be able to take the full context into consideration, or (and this is less time-consuming) these forms need to be checked manually after automatic PoS-tagging when creating gold standards. Alternatively, combined tags can be used (e.g. y ‘to his’ as P-PRO-G), but as indicated above, this further expansion of the tagset might complicate subsequent automatic parsing and any linguistic queries. A different solution to this type of problem is doing another round of editing when preprocessing, in this particular case to insert the missing y. But this again, is tedious and time-consuming and a job better suited for philologists than for syntacticians.

2.4 Opportunities

Addition of tags indicating prepositional and verbal inflection in Welsh opens up a new field of syntactic research. For example, it easily facilitates queries concerning all kinds of agreement phenomena. Since subject-verb agreement could be a diagnostic for unmarked V2 structures in Middle Welsh, this particular extra piece of information alone is invaluable in a diachronic study of the loss of V2. Adding person, number and gender features is furthermore crucial in coreference resolution and Information Structure. Finally, a detailed annotation scheme as the one proposed here for Welsh will not just create more research opportunities for syntacticians, but it will be equally beneficial for scholars of historical morphology and phonology. To facilitate future research even further, the annotated corpus should be available in a sustainable and adaptable format such as XML (TEI P5) (cf. among others, Komen (2013)).

3 Conclusion

This paper presents a good test case for annotating a historical corpus of a language with rich verbal and prepositional inflection. The main challenges in assembling this corpus lie in the availability of good diplomatic or critical text editions. Further collaboration with scholars specialised in the philological background producing these editions can help syntacticians to make the right decisions, both in terms of selecting the right texts and editions for the corpus, but also in preprocessing and tokenisation in particular. Essentially, Rissanen (1989)’s second problem of the risk over overgeneralisation (“God’s truth fallacy”), could be addressed with help from philologists (thereby also solving his first problem of the “philologist’s dilemma”).

Preprocessing the texts and expanding the training set takes time, but will yield better results in the end. Limiting the tagset will limit future research opportunities, because some features might be difficult to add in a later stage. A standardised way of expanding the tagset for “rich inflectional” languages is called for. The benefits of more research opportunities for scholars in various fields outweigh the initial, increased time-investment. It is possible to limit those extra efforts even further, for example, by using PoS-taggers that can easily deal with inflectional suffixes.

References