PoS-tagged Middle Welsh texts from Oxford, Jesus College MS. 119

Elena Parina, Raphael Sackmann and Marieke Meelen

The data result from the project 'Translations as language contact phenomena: studies in lexical, grammatical and stylistic interference in Middle Welsh religious texts' (led by Prof Erich Poppe, Philipps-Universität Marburg) 1 October 2015-30 September 2018, funded by the Fritz Thyssen Stiftung. The main publication from the project, currently under preparation, will be Parina & Poppe (fthc.).

The source data, transcriptions of the texts from Oxford, Jesus college MS. 119, also known as the *Book of the Anchorite of Llanddewi Brefi* or *Llyfr yr Ancr*, were obtained from the Cardiff project Rhyddiaith Gymraeg/ Welsh Prose 1300-1425 (with kind permission from Prof Sioned Davies and Dr Diana Luft).

The automatic part-of-speech tagging was completed by Dr Marieke Meelen (University of Cambridge). Manual corrections were done by Dr Elena Parina and Raphael Sackmann M.A.

These PoS-tagged and also syntactically annotated (chunk-parsed) files are also stored as part of the Parsed Historical Corpus of the Welsh Language (PARSHCWL http://lion.ling.cam.ac.uk/parshcwl/index.html), however we thought it useful to create and store separately the PoS-tagged files, since they have the simplest txt-format and offer data for different sorts of linguistic analysis, esp. using regular expressions.

The names of the files correspond to the following titles of the texts given in the *Rhyddiaith Gymraeg* project and to the pages in the manuscript:

File name	*title in Luft 2013*	
LlA-3v	*none*	
LlA-4v	*none*	
Lucidar	*Ystoria* ***Lucidar***	5r-69v
Marwolaeth	***Marwolaeth*** *Mair*	69v-77r
Ymborth	***Ymborth*** *yr Enaid*	78r-92v
Dewi	*Buchedd* ***Dewi***	93r-103v
Beuno	*Buchedd* ***Beuno***	104r-110r
Adrian	*Ystoria* ***Adrian*** *ac Ipotis*	111r-119r
Credo	***Credo*** *Athanasius*	119r-121r
Paddelw	***Pa ddelw*** *y dylai dyn gredu y Dduw*	121r-125r
Pader	*Pwyll y* ***Pader*** *o ddull Hu Sant*	125r-128r
Rhinweddau	***Rhinweddau*** *Gwrando Offeren + Rhinweddau Gweled Corff Crist*	128v-129r
Pawl	*Breuddwyd* ***Pawl***	129r-132v
Epistol	***Epistol*** *y Sul*	132v-134r
Luc1	***Rhybudd*** *Gabriel*	134r-134v
Ieuan1	***Efengyl*** *Ieuan*	134v-136r
Trindod	***Y Drindod*** *yn un Duw*	136r-137r
Gwlad	***Gwlad*** *Ieuan Fendigaid*	137v-143v

LlA-3v and LlA-4v are transcriptions of the first two pages of the manuscript written by its scribe; these feature the contents of the manuscript (3v) and the famous colophon with the reference to the manuscript's date, patron, and scribe as well as the prologue to *Lucidar* (4v).

The work stages were the following:

1. Preprocessing (March 2016) – Raphael Sackmann & Elena Parina

Special characters were normalized, such as <6> to <w>, in order to enhance searchability.
Tokens were divided and combined to facilitate further tagging.
Some emendations were made following editions of the texts or parallel readings from other manuscripts, they are marked with #.
Sentence boundaries were introduced (with <utt> markers), according to our understanding of Middle Welsh syntax. As subsequent work showed, this operation depends in a high degree on theoretical background and personal decisions and affects measurements of such features as word order.

2. POS Tagging (September 2016- March 2017)

Tagger: Marieke Meelen (see Meelen 2016)

Manual corrections: Raphael Sackmann & Elena Parina with advice from Erich Poppe

The tags are explained in Meelen (2016: 328-330) and are an expanded version of the tagset used in UPenn Historical corpora (https://www.ling.upenn.edu/histcorpora/).

The parser was first trained on the Middle Welsh Mabinogion corpus compiled by Meelen (2016) and then applied to three *Llyfr yr Ancr* texts. These texts were then manually corrected and included into the training corpus, which led to an improved performance of the tagger on the next set of the texts. The annotators aimed at consistency throughout the texts, but some effects of this multi-staged procedures might be reflected in the tagging. The texts were tagged and corrected in the following order:

1. Marwolaeth
2. Rhinweddau
3. Epistol
4. Pa ddelw
5. Pader
6. Pawl
7. Luc1
8. Trindod
9. Gwlad
10. Ieuan1
11. Credo
12. Lucidar
13. Adrian
14. Ymborth
15. Beuno
16. Dewi

An annotation log was kept to document all the decisions and changes and is available from Elena Parina at request.

Bibliography

Luft, Diana, Peter Wynn Thomas, and D. Mark Smith. 2013. Rhyddiaith Gymraeg 1300-1425. http://www.rhyddiaithganoloesol.caerdydd.ac.uk.

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

Parina, Elena, and Erich Poppe. fthc. *Translating devotion in medieval Wales: Studies in the texts and language of Llyfr Ancr Llanddewibrefi*. (with contributions by Meelen, Marieke, Sackmann, Raphael, and Scherschel, Ricarda)