Journal of Open Humanities Data

A Part-of-Speech Tagger for Older Scots: A spaCy-Based Model Using the CLAWS7 Tagset

COLLECTION: DIGITAL RESOURCES FOR THE LANGUAGES IN IRELAND AND BRITAIN

DATA PAPER

BETH BEATTIE

ubiquity press

ABSTRACT

This paper presents a part-of-speech (PoS) tagger developed for Older Scots. The tool was developed using spaCy machine learning models and trained on approximately 100,000 words of pre-tagged Older Scots literature. The tagger's underlying spaCy model uses the CLAWS7 PoS tagset instead of the Universal Dependencies tagset typically used in spaCy. The overall accuracy of the tagger is 89.83%, with performance constrained by the limited availability of Older Scots material and by considerable spelling variation. Despite these limitations, the tagger represents an important step forward in the computational treatment of Older Scots and provides a practical tool for researchers working with historical Scottish linguistic material. The spaCy model and tagging script are available through the JOHD Dataverse repository.

CORRESPONDING AUTHOR:
Beth Beattie
English Language and Linguistics, University of Glasgow, Glasgow, United Kingdom
Beth.Beattie@glasgow.ac.uk

KEYWORDS:
Older Scots; part-of-speech tagging; corpus linguistics; historical linguistics; natural language processing

TO CITE THIS ARTICLE:
Beattie, B. (2026). A Part-of-Speech Tagger for Older Scots: A spaCy-Based Model Using the CLAWS7 Tagset. *Journal of Open Humanities Data*, 12: 117, pp. 1–5. DOI: https://doi.org/10.5334/johd.611

(1) OVERVIEW

REPOSITORY LOCATION

The dataset is available at Dataverse: https://doi.org/10.7910/DVN/8VLSEK (Beattie (2026))

CONTEXT

This resource was created as part of a doctoral research project using corpus methods to explore the language of the Scottish and English Reformations. The corpus contains Early Modern English and Older Scots (OSc) texts[1] —related languages but different enough to require separate annotation approaches. While automated taggers are well-established for modern and well-resourced languages, including CLAWS for Early Modern English (Garside & Smith, 1997), significant challenges remain when developing tools for historical varieties and low-resource languages, particularly where training data is limited and there is a large amount of linguistic variation.

Recent projects have developed annotation tools for other low-resource languages. Lameris and Stymne (2021) found that a tagger trained only on English performed poorly when applied to Scots without any Scots-specific training (42–56% accuracy), showing that Scots cannot simply be treated as a variant of English. Adding a small amount of manually annotated Modern Scots data improved accuracy to 67.4%, though this remained limited by the quality of available word embeddings—numerical representations of word meaning used by the model—and the coarse granularity of the Universal Part of Speech (UPOS) tagset used (Petrov et al., 2012). Millour et al. (2024) applied a similar approach to Corsican, improving an Italian-trained tagger from 62.84% to 93.38% accuracy through fine-tuning on manually annotated Corsican material, with spelling variation and rare tags being the primary challenges. OSc faces similar challenges, with the added difficulty of limited surviving material typical of historical languages.

No PoS tagger has previously been trained specifically on OSc using machine learning methods. While some existing work has applied rule-based approaches to tagging OSc material—Gotthard's PhD research involved tagging OSc material using Python scripts alongside dictionaries from the Linguistic Atlas of Older Scots and the Penn Treebank (Gotthard, 2023, pp. 38–40)—such methods are time-consuming, require exhaustive manual specification of linguistic rules, and rely on the quality of the dictionaries to capture orthographic variation. The tagger described here addresses this gap: a spaCy-based model built using the spaCy natural language processing library trained on pre-tagged OSc literary texts (Bushnell, 2021), achieving 89.83% accuracy using pre-trained weights from a high-resource English model. The tagger uses the CLAWS7 tagset (UCREL, 1996), which offers considerably finer morphosyntactic granularity than the UPOS tagset.

(2) METHOD

STEPS

The tagger was built using spaCy (Honnibal et al., 2020), an open-source Python natural language processing library that supports tasks including tokenisation, lemmatisation, and PoS tagging, allows custom taggers to be trained. spaCy's neural tagging approach was suited to OSc because it predicts tag sequences by representing words as token embeddings and incorporating character-level information to handle the orthographic variation present in OSc.

The spaCy model for the tagger was trained using two tagsets: UPOS, required by spaCy, and CLAWS7, the primary output tagset. CLAWS7 contains 152 tags which can be grouped into larger macro-categories (for example, all noun tags begin with N, proper nouns NP, etc.), allowing for both broad and narrow searches of tagged data. The training data was a ca.100,000-word corpus of pre-tagged OSc literary texts compiled by Bushnell (2021), covering OSc literature written before 1600. The files were pre-processed to remove extraneous etymological and semantic information, retaining the original word forms, lemmas, and CLAWS7 tags. A conversion script was used to map CLAWS7 tags to their UPOS equivalents, and the files were converted from

1 Periodisation of Scots and English follows the conventions outlined in Kopaczyk (2013), with Early Modern English and Older Scots in this paper referring to the language varieties used between c.1450 and c.1700.

JSON to the .spacy format required for training. The corpus was divided into three sets: 'train' (two thirds of the data, used to train the model), 'dev' (used by the model to check itself during training), and 'test' (used for formal accuracy evaluation afterwards).

Three model variants were compared, differing in their use of pre-trained resources: training data only; pre-trained word vectors from the English-language spaCy model en_core_web_sm; and pre-trained weights from en_core_web_lg. After initial training on the Bushnell corpus, an iterative process was used, whereby individual OSc texts outside the original training set were tagged, reviewed, and added to the training data. This allowed the tagger to develop familiarity with vocabulary outside of the literary training data, broadening the tagger's applicability to other OSc research.

SAMPLING STRATEGY

The choice of training data was constrained by the availability of existing annotated OSc material. The corpus of OSc literature selected as the primary training source was chosen because the size of the corpus—approximately 100,000 words—provided a substantial body of OSc material for the model to work with, and the texts were already tagged using the CLAWS7 tagset. While normalised spellings of the training data were available in the dataset, these were normalised according to English orthographic conventions. Therefore, original OSc spellings were retained in the training data rather than normalised forms to provide the model with as broad a range of OSc orthographic forms as possible.

QUALITY CONTROL

Model accuracy was evaluated using two methods. First, spaCy's built-in evaluate command was used to produce an overall accuracy score for each of the three model variants. A second bespoke evaluation script was also written to compare model output against the 'test' set, generating precision (proportion of positive predictions correctly identified), recall (proportion of all positive predictions identified), and F1-scores (harmonic mean of precision and recall) for each of the 118 tags present in the corpus. This allowed for a more granular assessment of performance across tags, distinguishing high-frequency tags (F1 over 0.74) from low-frequency tags, whose performance was more variable. On the basis of these evaluations, the model using pre-trained weights was determined to be the most accurate, with an overall accuracy of 89.83%.

A limitation of the CLAWS7 tagset is the absence of a dedicated punctuation tag, which resulted in punctuation tokens being assigned arbitrary tags during the tagging process. This was resolved post-hoc using a Python script to replace incorrectly assigned punctuation tags with the correct punctuation symbol. Manual review of tagger outputs during the iterative tagging process also identified recurring error types, particularly around the disambiguation of OSc morphosyntactic features such as the <-is> suffix, which can represent plural nouns, verb endings, and genitive markers. These errors were documented and will inform future development of the tagger.

(3) DATASET DESCRIPTION

REPOSITORY NAME

Dataverse

OBJECT NAME

'A Part-of-Speech Tagger for Older Scots: A spaCy-Based Model Using the CLAWS7 Tagset'

FORMAT NAMES AND VERSIONS

Training data: .spacy

Tagging script: .py

Accepted input: .txt

Outputs: .csv and .xml

CREATION DATES

2024-08-10–2026-06-24

DATASET CREATORS

Beth Beattie, University of Glasgow

LANGUAGE

Older Scots

LICENSE

CC0 1.0

PUBLICATION DATE

2026-06-24

(4) REUSE POTENTIAL

The most immediate reuse potential of this tagger is its applicability to other OSc corpora. Researchers working with historical Scottish texts can apply this tagger directly to OSc texts, facilitating larger-scale linguistic analysis. Beyond direct application, the model and training data can support further development, whether through fine-tuning on additional OSc texts or as a methodological reference for tagger development in other historical or low-resource languages.

There are limitations that should be acknowledged, however. The tagger was trained on literary and religious OSc texts, and its performance on other genres—particularly legal and administrative writing—has not been evaluated. Performance is also lower for low-frequency tags, and the disambiguation of morphosyntactically ambiguous OSc features remains a known source of error. Researchers applying the tagger to new material are encouraged to evaluate performance on a representative sample before relying on outputs for detailed analysis.

AI DECLARATION

No AI was used in the creation of this manuscript outside of basic formatting and proofreading.

ACKNOWLEDGEMENTS

Many thanks to Dr Megan Bushnell, Oxford Text Archive, for supplying the training data for the model.

FUNDING STATEMENT

This work was completed as a part of PhD research funded by the College of Arts and Humanities at the University of Glasgow.

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR CONTRIBUTIONS

Beth Beattie: Conceptualisation, Data curation, Investigation, Methodology, Software, Formal Analysis, Validation, Writing – original draft, Writing – review & editing.

AUTHOR AFFILIATIONS

Beth Beattie orcid.org/0009-0002-4649-5195
English Language and Linguistics, University of Glasgow, Glasgow, United Kingdom

REFERENCES

Beattie, B. (2026). *A Part-of-Speech Tagger for Older Scots: A spaCy-Based Model Using the CLAWS7 Tagset* (Version 1) [Data set]. Harvard Dataverse. https://doi.org/10.7910/DVN/8VLSEK

Bushnell, M. (2021). *Equivalency, page design, and corpus linguistics: An interdisciplinary approach to Gavin Douglas's 'Eneados'* [PhD Thesis, University of Oxford]. https://ora.ox.ac.uk/objects/uuid:1ee08a4e-8a00-4641-b368-1d568b97ac31

Garside, R., & **Smith, N.** (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. N. Leech, & A. McEnery (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora* (pp. 102–121). Harlow: Longman. https://doi.org/10.4324/9781315841366

Gotthard, L. (2023). *Syntactic change during the anglicisation of Scots: Insights from the Parsed Corpus of Scottish Correspondence* [PhD Thesis, The University of Edinburgh]. https://doi.org/10.7488/era/3154

Honnibal, M., Montani, I., Van Landeghem, S., & **Boyd, A.** (2020). *explosion/spaCy: Industrial-strength natural language processing in Python* (v3.7.2) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.1212303

Kopaczyk, J. (2013). Rethinking the traditional periodisation of the Scots language. In J. Cruickshank & R. M. Millar (Eds.), *After the Storm: Papers from the Forum for Research on the Languages of Scotland and Ulster triennial meeting, Aberdeen 2012* (pp. 233–260). Aberdeen: Forum for Research on the Languages of Scotland and Ulster.

Lameris, H., & **Stymne, S.** (2021). Whit's the Richt Pairt o Speech: PoS tagging for Scots. In M. Zampieri, P. Nakov, N. Ljubešić, J. Tiedemann, Y. Scherrer, & T. Jauhiainen (Eds.), *Proceedings of the 8th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects* (pp. 39–48), Kiyv, Ukraine. https://aclanthology.org/2021.vardial-1.0/

Millour, A., Brasile, L., Ghia, A., & **Kevers, L.** (2024). Agettivu, Aggitivu o Aghjettivu? POS Tagging Corsican Dialects. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 600–608). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.52/

Petrov, S., Das, D., & **McDonald, R.** (2012). A Universal Part-of-Speech Tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2089–2096). Istanbul, Turkey: European Language Resources Association (ELRA). https://doi.org/10.63317/3pjyez8kmkhz

UCREL. (1996). *CLAWS7 Tagset.* University of Lancaster. UCREL NLP Group. Retrieved from https://ucrel.lancs.ac.uk/claws7tags.html (last accessed: 24 July 2026).

TO CITE THIS ARTICLE:
Beattie, B. (2026). A Part-of-Speech Tagger for Older Scots: A spaCy-Based Model Using the CLAWS7 Tagset. *Journal of Open Humanities Data*, 12: 117, pp. 1–5. DOI: https://doi.org/10.5334/johd.611

Submitted: 26 June 2026
Accepted: 30 July 2026
Published: 27 August 2026

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.