Journal of Open Humanities Data

Gavin Douglas's *Eneados*: An Older Scots Text Formatted as Corpus Data

COLLECTION: DIGITAL RESOURCES FOR THE LANGUAGES IN IRELAND AND BRITAIN

DATA PAPER

MEGAN BUSHNELL

ubiquity press

ABSTRACT

This data paper describes the 'corpus-based apparatus' created for a doctoral project that produced a scalable reading of Gavin Douglas's *Eneados* – a sixteenth-century translation of the *Aeneid*, written in Older Scots. This resource is deposited in the Oxford Text Archive, a national repository for literary and linguistic data, and includes a fully digitised version of the *Eneados* and its Latin source texts, Virgil's *Aeneid* and Maffeo Vegio's *Supplement*, in both TEI-XML and plain text formats, along with XSLT and sed scripts involved in the development of these files. All texts were extracted from preexisting digital transcriptions and the Latin texts were manually edited to reflect the orthography and materiality of Douglas's source, an early printed book of Virgil's works. The TEI-XML files are heavily annotated with linguistic and structural information, much of which has the potential to be reused for the development of tools for processing Older Scots and potentially even be of use for Modern Scots – an under-resourced, national language of Scotland.

CORRESPONDING AUTHOR:
Megan Bushnell
Faculty of Linguistics, Philology and Phonetics, University of Oxford, Oxford, UK
megan.bushnell@ling-phil.ox.ac.uk

KEYWORDS:
Gavin Douglas; Eneados; Older Scots; translation; classical reception; linguistic data

TO CITE THIS ARTICLE:
Bushnell, M. (2026). Gavin Douglas's *Eneados*: An Older Scots Text Formatted as Corpus Data. *Journal of Open Humanities Data*, 12: 147, pp. 1–6. DOI: https://doi.org/10.5334/johd.648

(1) OVERVIEW

REPOSITORY LOCATION

https://hdl.handle.net/20.500.14106/2580

https://doi.org/10.5281/zenodo.23568828

CONTEXT

Gavin Douglas's *Eneados* (1513) is a text of special linguistic interest. Not only is it the first full, direct translation of the *Aeneid* produced in Britain, it is one of the first texts to identify Scots as a language separate from English – though not, of course, the first to be written in Scots, which has a long, distinguished literary history. In addition, it has been the subject of intense philological scrutiny: the early philologists William L'Isle (1569–1637) and Franciscus Junius the Younger (1591–1677) used this text to reconstruct Anglo-Saxon and elucidate Middle English, and Thomas Ruddiman (1675–1757) appended a glossary to his 1710 edition of the *Eneados*, which became the basis for John Jamieson's (1759–1838) *Etymological Dictionary of the Scottish Language* (1808; Watt, 1920).

The resource described here was produced by a doctoral project at the University of Oxford that performed a scalable reading (that is to say, a combination of close and distant readings) of the *Eneados* using corpus linguistic methods and tools (Bushnell, 2021a). For this reason, this resource is described as a 'corpus-based apparatus', in that it is informed by corpus linguistic principles, but is not a corpus in its own right.

This resource has since been used in a multitude of papers regarding Gavin Douglas, including one on his response to his source text's layout (Bushnell, 2021b), one on his use of repetition in translation (Bushnell, 2025a), and one on the nature of his nationalism (Bushnell, 2025b). It has also been used in a small project with the *Oxford English Dictionary* to create a pilot etymological tagger for Older Scots (Bushnell, 2019), and has been used in the work of Beth Beattie, who used the part-of-speech annotation to create a part-of-speech tagger for Older Scots (Beattie, 2026).

(2) METHOD

The sources for the digital files created for this project include Coldwell's edition of the *Eneados* (Coldwell, 1957–60), Greenough & Kittredge's edition of the *Aeneid* (Greenough & Kittredge, 1897–1902), and Brinton's edition of the *Supplement* (Brinton, 1930). These editions were selected primarily because they are all already digitised and easily accessible online, though there were other factors, including manuscript witnesses (Bushnell, 2021a).

STEPS

With the base data obtained, the following workflow was applied to create the final resource:

1. The digital transcriptions of the *Eneados*, *Aeneid*, and *Supplement* from the sources listed above were put into ASCII plain text files. Texts of the *Aeneid* and *Supplement* were manually adapted so that their content, layout, lexis, orthography, and punctuation all matched those of Douglas's source, Ascensius's Paris edition of the complete works of Virgil (1501), as established by Priscilla Bawcutt (1973).
2. The ASCII files were transformed into Unicode to allow for the use of special characters – specifically, the yogh (ȝ) character, which is a common feature of Older Scots. These Unicode files are the plain text versions of the texts and have no line numbers.
3. Line numbers and other structural information, including chapter numbers, segmentation of the Latin text, and page breaks in Ascensius's edition (1501) were added to the Unicode files.
4. Two different file versions were created:
    a. The Scots and Latin files were aligned manually by line. The texts were divided into translation units, which contained the smallest number of whole lines (barring interference from layout) of the source and translation that correspond to one

another so that the source is completely translated (barring intentional elision), and the translation is completely accounted for (barring intentional deviation). The aligned files were originally designed to be used with *AntPConc* (Anthony, 2014), which expects equivalent content across two text files to be present on the same line in their respective files. Later, in the XML format, these translation units were represented by <ab> or 'anonymous block' elements, and given a unique ID, so that each translation unit in the source text could be matched to equivalent material in the translation.

   b. Pragmatic annotations indicating boundaries of narrative, speech, and reported speech, as well as the name, gender, status, and nationality of each speaker, were added to the Unicode files.

5. The Unicode files from 4b were transformed into TEI-XML. Some light customisations were used; these are all declared in the TEI header of each file.
6. The TEI-XML files were tokenised so that each word was contained in its own <w> element. This tokenisation was automated using an XSLT (eXtensible Stylesheet Language Transformations) script. XSLT can transform an XML document into other formats; it can also manipulate the structure within an XML document, making it useful for tasks like tokenisation and annotation. The XSLT script for tokenisation has been deposited with the data. The Latin files were not developed further than this step.
7. The Scots files were further enhanced. Two different file versions were created:
   a. The Scots TEI-XML files were all normalised (and, where necessary, modernised) to enable something akin to lemmatised search and part-of-speech and semantic tagging using automated tools designed to be used with Modern English. This process was semi-automated. A list of every type in the body of the *Eneados* was generated using *AntConc*'s wordlist function. Each type was manually assigned a normalised form, and, in some cases, a modern equivalent, with reference to the *Oxford English Dictionary* and *Dictionary of the Scots Language*. This generated a normalised lexicon, which is part of the dataset. Using sed – a UNIX stream editor that can transform texts – each 'word' element (<w>) in the XML files was matched with its normalised form according to the word list. The original, normalised, and modernised forms were then inserted into the <w> element as attributes (@orig, @norm, @mod). The normalisation then was checked and revised manually.
   b. Normalised versions of all the Scots files were generated using an XSLT script, which replaced the value of all <w> elements with the value of @norm, or, where necessary, @mod attributes. This XSLT script has been deposited with the data.
8. Part-of-speech and semantic tagging was applied to the Scots files from 7b using the *USAS tagger* (Archer et al., 2003). The tagging was then checked and revised manually and all unidentified tags (tag value Z99) resolved.
9. The Scots files from steps 6 and 8 were merged into one file containing the original spelling, with normalised and modernised annotations, and part-of-speech and semantic tagging (preserved as @pos and @sem attributes). This was done using an XSLT script, which has been deposited with the data.
10. The Scots files were then enhanced with information from the 4a files manually, so that alignment information was encoded into the TEI-XML files as <ab @type = "translation_unit"> elements, with numeric identifiers. These are the final versions of the encoded Scots files.

QUALITY CONTROL

Throughout the creation of these files, the normalisation, and part-of-speech and semantic tagging were manually evaluated and revised to ensure maximum accuracy. However, no large scale benchmarking has been undertaken at present.

(3) DATASET DESCRIPTION

REPOSITORY NAME

The Oxford Text Archive

Zenodo

OBJECT NAME

Gavin Douglas's *Eneados*, annotated, tagged, and aligned with its source texts Virgil's *Aeneid* and Maffeo Vegio's *Supplement*.

FORMAT NAMES AND VERSIONS

The resource consists of digital files, schema files, and scripts involved in the resource's creation. The digital files consist of 39 encoded files in TEI-XML and 39 of the same files in Unicode plain text. For a full breakdown of the word count within each file see the author's DPhil thesis (Bushnell, 2021a, Table 2). The schema file is in RNC, with documentation in HTML. The scripts are predominantly in XSLT, along with one SH file (shell script) for applying normalisation using sed, and a CSV file containing the normalised lexicon.

CREATION DATES

2015-10-01–2021-05-01

DATASET CREATORS

Megan Bushnell, University of Oxford. Conceptualisation and data curation.

Martin Wynne, University of Oxford. Software (specifically the sed script) and supervision.

LANGUAGE

Raw data: Older Scots and Latin.

Metadata and annotation: English.

LICENSE

CC BY-NC-SA 4.0

PUBLICATION DATE

2025-06-26

(4) REUSE POTENTIAL

This dataset has already been used extensively to create a computational profile of Gavin Douglas's language use – especially his practice in translation, his use of repetitions, and his response to facets of layout in his source text (Bushnell, 2021a, 2021b, 2025a, 2025b). More can still be done on this topic – for example, stylometric analysis of Douglas's language to computationally refute his contested authorship of *King Hart* (Preston, 1959), extraction and analysis of Douglas's translation pairs, and investigation of how he uses 'double translation' in his works (Bawcutt, 1976, p. 116). This dataset could also be combined with digitised texts by other medieval Scots authors to create a medieval Scots literary corpus. While corpora of this sort do already exist, like the Aberdeen Corpus of Older Scots (Discry, 2011), none of these have been annotated to the extent that this resource has been.

While this dataset is focused on the language of a specific author in a specific text, it has the potential to be more widely applicable. In particular, the linguistic annotation could be extracted and used as training data to create Older Scots specific tools. Beth Beattie has already used this data to help develop a part-of-speech tagger for Older Scots that she has applied in her research (Beattie, 2026). Perhaps, in future, a semantic tagger could also be created, or even a language embedding model specifically for Older Scots. These in turn could be of use to

related languages, such as Modern Scots, and Middle English. This data can also be used as the foundation of a digital edition of the *Eneados*. Preliminary work has been undertaken on this account, with the creation of a static parallel reader called *Aeneas* (Bushnell, 2026).

AI DECLARATION

AI was not employed at any stage of the manuscript drafting.

AI was not employed as part of data acquisition or processing.

ACKNOWLEDGEMENTS

Special thanks to Martin Wynne and Sally Mapstone, who supervised this project.

FUNDING STATEMENT

This project was funded in part by Mount Holyoke College's Frances Mary Hazen Fellowship, and by the University of Oxford's Vice-Chancellor's Fund.

COMPETING INTERESTS

The Author is a guest editor for this special collection on Digital Resources for the Languages in Ireland and Britain, but had no involvement in any editorial processes related to the handling of this submission.

AUTHOR CONTRIBUTIONS

Megan Bushnell – Conceptualisation, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, validation, visualisation, writing – original draft, writing – review & editing.

AUTHOR AFFILIATIONS


Megan Bushnell orcid.org/0000-0002-2819-0585
Faculty of Linguistics, Philology and Phonetics, University of Oxford, Oxford, UK

REFERENCES

Anthony, L. (2014). *AntPConc* (Version 1.1.0) [Computer software]. Waseda University. https://www.laurenceanthony.net/software/antpconc/

Archer, D., McEnery, T., Rayson, P., & **Hardie, A.** (2003). Developing an automated semantic analysis system for Early Modern English. In D. Archer (Ed.), *Proceedings of the Corpus Linguistics 2003 Conference*. UCREL, Computing Department, University of Lancaster.

Bawcutt, P. (1973). Gavin Douglas and the Text of Virgil. *Edinburgh Bibliographical Transactions*, *4*(6), 211–231.

Bawcutt, P. (1976). *Gavin Douglas: A Critical Study*. Edinburgh University Press.

Beattie, B. (2026). *A Part-of-Speech Tagger for Older Scots: A spaCy-Based Model Using the CLAWS7 Tagset* [Dataset]. Harvard Dataverse. https://doi.org/10.5334/johd.611

Bushnell, M. (2019, October 14). *Using the OED to investigate the implications of Douglas's lexical choices in the Eneados* [Webinar]. YouTube. https://www.youtube.com/watch?v=0H-MBK9O4qk

Bushnell, M. (2021a). *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

Bushnell, M. (2021b). Reconstructing Gavin Douglas's translation practice in the *Eneados* using a corpus linguistic-based method. *DHBenelux*, *3*, 13–38. https://doi.org/10.17613/p3z7-4c05

Bushnell, M. (2025a). Creating a literary *koine*: how Gavin Douglas translates repetition in the *Eneados*. In D. Bertagnolli & A. Zironi (Eds.), *Linguistic Fragmentation and Cultural Inclusion in the Middle Ages: Translation, Plurilingualism, Multilingualism* (Medieval Translator, Vol. 22, pp. 183–208). Brepols. https://doi.org/10.1484/M.TMT-EB.5.142362

Bushnell, M. (2025b). What's in a name? Douglas's depiction of nationhood in the *Eneados*. *Journal of the Northern Renaissance*, *15*. https://jnr2.hcommons.org/2025/8648/

Bushnell, M. (2026). *Aeneas: Static Parallel Reader*. https://mebushnell.github.io/Aeneas/

Discry, C.-H. (2011). *The Aberdeen Corpus of Older Scots (ACOS)* [Data set]. University of Aberdeen. https://hdl.handle.net/20.500.14106/2583

Douglas, G. (1957–60). *Virgil's 'Aeneid' translated into Scottish verse by Gavin Douglas* (D.F.C. Coldwell, Ed.; 4 vols.). Scottish Text Society. https://www.proquest.com/docview/2148064579

Preston, P. (1959). Did Gavin Douglas Write *King Hart*? *Medium Aevum*, *28*. https://doi.org/10.2307/43626771

Vegio, M. (1930). *Maphaeus Vegius and his thirteenth book of the 'Aeneid': A chapter on Virgil in the Renaissance* (A.C. Brinton, Ed.). Stanford University Press. https://virgil.org/supplementa/vegio-twyne.htm

Virgil. (1501). *Opera cum variorum commentariis* [Shelfmark: Ink 4. D 7672, UB Freiburg, Freiburg, Germany]. Ascensius & Jean Petit. https://dl.ub.uni-freiburg.de/diglit/vergilius1500

Virgil. (1897–1902). *The Bucolics, Aeneid, and Georgics of Virgil* (J.B. Greenough & G.L. Kittredge, Eds.; 2 vols.). Ginn. https://data.perseus.org/texts/urn:cts:latinLit:phi0690.phi003.perseus-lat1

Watt, L.M. (1920). *Douglas's Aeneid*. Cambridge University Press.

TO CITE THIS ARTICLE:
Bushnell, M. (2026). Gavin Douglas's *Eneados*: An Older Scots Text Formatted as Corpus Data. *Journal of Open Humanities Data*, 12: 147, pp. 1–6. DOI: https://doi.org/10.5334/johd.648

Submitted: 02 September 2026
Accepted: 07 October 2026
Published: 13 November 2026

COPYRIGHT:
© 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See https://creativecommons.org/licenses/by/4.0/.

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