Ansund: Using Machine Learning to Develop a New, Exhaustive, Open Access Corpus of Old English

Old English is the earliest written phase of English, and occupies a particularly important place in the emergence of European vernacular literacy, by virtue of the thousands of texts and millions of words of it that survive. The standard electronic repository of Old English texts is the Dictionary of Old English Corpus (DOEC). The current (2009) version is disseminated online, via subscription. DOEC is the bedrock of all philological and corpus linguistic research on Old English.

Yet DOEC has significant limitations. Its texts are overwhelmingly based on published editions, so to an unknowable extent encode editorial normalisations at the expense of the scribal variants that are most important key for historical linguists. DOEC also includes only one version of most works, therefore representing only one quarter to one third of surviving Old English. Since copyright still resides in the editions underpinning DOEC, datasets developed from it cannot be made available online. Coupled with DOEC's subscription-based financial model, this is unavoidably an obstacle to open science. Ansund, which takes its name from an Old English word meaning 'whole, healthy [in body]', will remedy these corpus problems.

It will do so via HTR. Importantly, HTR fundamentally changes the nature of the questions we can ask. The conventions hitherto used for transcribing Old English (e.g. printing wyn as 'w', suppressing original punctuation marks and normalising word division) largely derive from the exigencies of the printing press and the need of students for minimal unfamiliarity. Yet, digital dissemination allows for the reproduction of an almost infinite array of letterforms, which a machine can transcribe far more accurately than a human, limited in patience and pre-trained to ignore many relevant features (volubly demonstrated by Stutzmann 2020 for word division). These affordances will allow us to ask many hithertho unanswerable questions, such as what acute accents mean, what patterns underpin Old English word division, and how questions were punctuated in Old English.

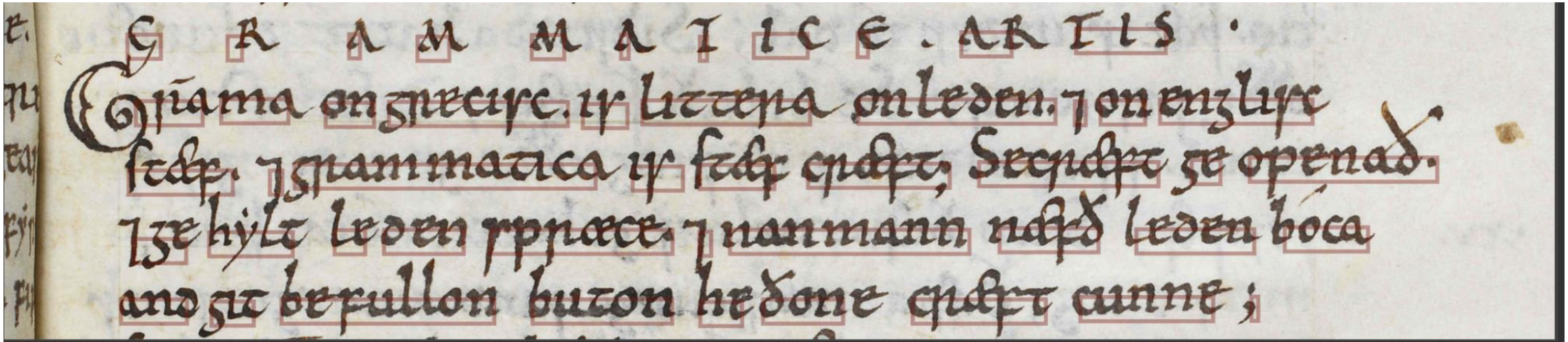
We are currently in the pilot phase of Ansund. We have transcribed samples from 25 manuscripts, ranging in date from the ninth to the twelfth centuries. A Transkribus model trained on these performs at 98% accuracy. We are now finalising transcription conventions (working with Tarrin Wills from MUFI and font designer Peter Baker, the developer of Junicode) and making key methodological decisions about how and if to record word division, glosses and corrections, prior to revising the transcriptions to these new conventions and retraining the model. We also wish to test whether noisy HTR output needs correction to be usable in Old English studies (Eder 2013 showed data could be degraded by 20% and still yield meaningful stylometric results). With this done, we will host a colloquium in late 2024 to determine with colleagues and library partners the optimal way of bringing Ansund to fruition, whether through grant funding or field-wide collaboration.

References

Healey, A. diP., Wilkin, J. P. and Xiang, X. 2009. *Dictionary of Old English Web Corpus*. <https://tapor.library.utoronto.ca/doecorpus/>

Eder, Maciej. 2013. Mind Your Corpus: Systematic Errors in Authorship Attribution. *Literary and Linguistic Computing* 28. 603-14.

Stutzmann, Dominique. 2020. Words as Graphic and Linguistic Structures: Word Spacing in Psalm 101 *Domine exaudi orationem mean* (Eleventh-Fifteenth Centuries). In Victoria Turner & Vincent Debais, eds, *Les Mots au Moyen Âge/Words in the Middle Ages*. Turnhout: Brepols. 21–57.



1-10 G R A M M A T I C E . A R T I S ·

1-11 Gꞃāma on ᵹꞃeciſc. iſ litteꞃa onleꝺen . ⁊ on enᵹliſc

1-12 ſtæꝼ . ⁊ ᵹꞃammaꞇica iſ ſtæꝼ cꞃæꝼꞇ ; Secꞃæꝼꞇ ᵹe openað .

1-13 ⁊ ᵹe hẏlt leꝺen ſpꞃæce . ⁊ nan mann næꝼð leꝺen bóca

1-14 anꝺ ᵹit beꝼullon buꞇon heꝺone cꞃæꝼꞇ cunne ;

Mark Faulkner

BIOGRAPHY

I am considered a world expert on the transition from Old to Middle English, thanks to a long series of articles published over the last ten years. My *New Literary History of the Long Twelfth Century*, 'news' and 'a game changer' according to reviewers, came out in 2022. Absent grammars or handbooks of Transitional English, this research had to proceed from the ground up. Since 2017, I have consequently begun developing, with €370k+ of IRC and Trinity College Dublin funding and Research Boost funding, a new subdiscipline of historical linguistics, 'corpus philology', which combines insights from corpus linguistics and traditional philology, to allow the contextualisation of the language of individual texts within the entirety of surviving medieval writing, using datasets enormous by comparison with the norms of the discipline. While my background is in English Literature, as a philologist I have also published and taught extensively in Book History and Linguistics, and indeed across the whole sweep of Medieval Studies, increasingly in collaboration with computer scientists.

CURRENT POSITION

2016- Ussher Assistant Professor in Medieval Literature, School of English, Trinity College Dublin, Ireland.

MAJOR INSTITUTIONAL RESPONSIBILITIES

2021- Lead Developer and Director, Trinity Centre for the Book, Trinity College Dublin

2016-2022 Lead Developer and Director, M. Phil in Medieval Studies, Trinity College Dublin, Ireland.

RECENT FUNDING

2023 Trinity Long Room Hub Research Incentive Scheme, 'Ansund: Using Machine Learning to Develop a New, Exhaustive, Open Access Corpus of Old English' (€5,000)

2023 Irish Research Council Collaboration Bursary (€2,000) [awarded since shortlisted for Irish Research Council Consolidator Laureate Award, 'Dating Medieval Texts']

2022-2024 Irish Research Council Coalesce Scheme, 'Searobend: Linked Metadata for English Language Texts, 1000-1300' (€219,225)

KEYNOTE LECTURES

05-24 'The Twain Shall Meet: Rethinking the Relationship between 'Old' and 'Middle' English', 13th International Conference on Middle English, Málaga

07-23 'Towards Medieval Big Data: corpora, metadata and methodologies for early English', 22nd Conference on English Historical Linguistics, Sheffield

12-21 'Corpus Philology: Towards Automated Linguistic Profiling of Old English Texts', 43rd Symposium on Old English, Middle English and Historical Linguistics in the Low Countries, Leiden

MAJOR PUBLICATIONS

M. Faulkner, *A Critical Anthology of Twelfth-Century English: Writing the Vernacular in the Transitional Period* (York: Arc Humanities Press, 2025) [c. 110k words, delivery date 09-24]

E. Bonapfel, M. Faulkner, J. Guttierez & J. Leonard, eds., *The History of Punctuation in English Literature* ed., 3 vols. (Cambridge, Cambridge University Press, 2025) [c. 1m words, delivery date 04-24]

M. Faulkner, *A New Literary History of the Long Twelfth Century: Language and Literature between Old and Middle English* (Cambridge: Cambridge University Press, 2022).

Winner, International Society for the Study of Early Medieval England Best First Monograph Prize (2023)

Major reviews: *Review of English Studies* 74 (2023), 354-5 ('a milestone in the study of early English … a formidable achievement'); *Speculum* 98 (2023), 868-9 ('far and away the best study of the period to date').

ELISABETTA MAGNANTI

University of Vienna, Department of History
Universitätsring 1, 1010 Vienna
+33 7 79 21 82 98 elisabetta.magnanti@univie.ac.at

EMPLOYMENT

Mar 2021 – Present **University Assistant (Prae-doc) in Digital Humanities | University of Vienna**
Department of History, Faculty of Historical and Cultural Studies

Nov 2019 – Feb 2021 **Project Officer | The French National Centre for Scientific Research (CNRS), Paris**
Department of Science and Information Technology (DIST)
Present-day *Open Research Data Department (DDOR)*

EDUCATION

2021 – Present **PhD in History | University of Vienna**
Department of History, Faculty of Historical and Cultural Studies

2018 – 2020 **MA in Germanic Philology | University of Siena**
Department of Philology and Literary Criticism

2017 – 2018 **MSc in Digital Humanities | University of Siena**
Department of Information Engineering and Mathematical Sciences
Department of Philology and Literary Criticism

FURTHER EDUCATION

June 2020 **Machine Learning for Big Data and Text Processing |Massachusetts Institute of Technology**
Concentrations: Deep Learning, Neural Networks, Artificial Intelligence

July 2019 **Digital Humanities Summer School |University of Oxford**
Strand: *Humanities Data: Case Studies and Approaches*

RECENT PUBLICATIONS

'Handwritten Text Recognition in Medieval Germanic Manuscripts: The Peterborough Chronicle as a Case Study', *Filologia Germanica – Germanic Philology* 14 (2022), 193–216.

Cerretini, G., Magnanti, E., and Rosselli Del Turco, R., 'TEI as Data', in *TEI 2022 Conference Book*, edited by James Cummings, 111–13, 2022.

INVITED LECTURES and CONFERENCE PRESENTATIONS

01 Dec 2023 **Using AI to Transcribe Trinity's Manuscripts**
with Mark Faulkner (TCD)
Trinity College Dublin (Ireland) | International Symposium 'The Many Lives of Medieval Manuscripts'

05 July 2023 **New Tools and Methods in Digital Medieval Studies: A Round Table Discussion**
with Stewart J. Brookes (Oxford), Aaron Macks (Harvard), Jan Odstrčilík (ÖAW), and Dot Porter (Pennsylvania)
Leeds International Medieval Congress (IMC) | University of Leeds (United Kingdom)

01 June 2023 **Methods and Tools for Digital Philology**
University of Pisa (Italy) | Summer School 'Digital Tools for Humanists: Working on History'

06 Dec 2022 **Using Machine Learning to Generate Big Data for Book History**
with Mark Faulkner
Trinity College Dublin (Ireland) | Lecture series 'Manuscript Book and Print Culture'