

racking the code

Challenges of Encoding Medieval Manuscripts

ofie oors (FWO PhD, University of Antwerp)

outer averals (FWO Postdoc, University of Antwerp)

Constrained

- Middle Ages: scribes making copies of copies (of copies...) by hand
- Each manuscript copy: unique variants
- Material Philology; research into how copyists manipulate a text (Van Dalen-Oskam 2012; Kestemont 2018; Haverals & Kestemont *forthcoming*)



- Little research on **underlying mechanisms** in copying process
- Influence of **text form** (such as rhyme scheme, stanza form, and text structure)
- Evidence from **psycholinguistics**: formal features = constraining character (Rubin 1995)
- Slowly, more **computational** research on this topic is shaping up (Thaisen 2014)

MEMORY

IN ORAL

TRADITIONS

*The Cognitive
Psychology
of Epic,
Ballads, and
Counting-out
Rhymes*



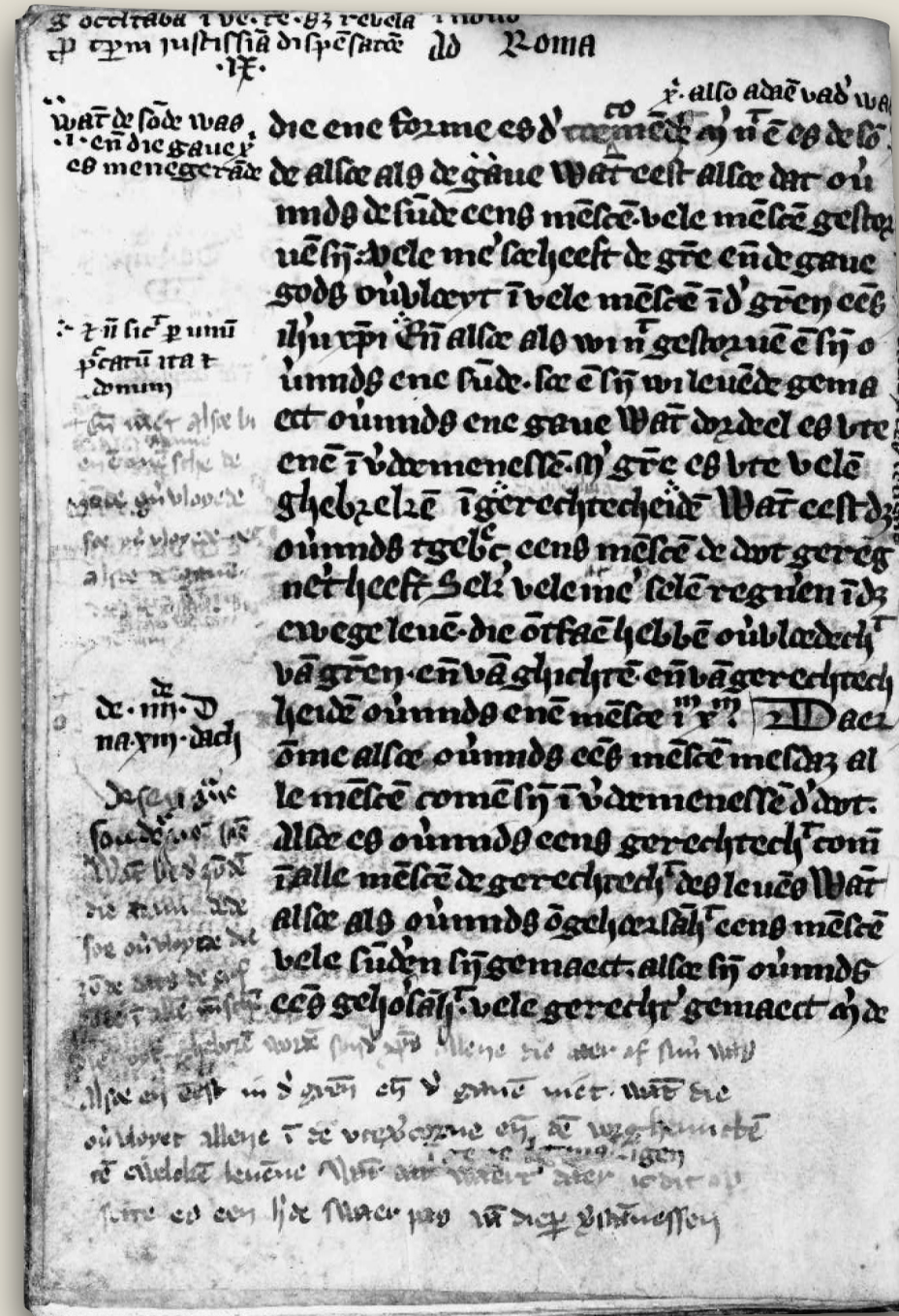
DAVID C. RUBIN

- Computational investigation of the transmission of *Martijn Trilogy* by Jacob van Maerlant
- **Computational alignment** of 17 different manuscripts, fragments, and prints before 1500
- **Empirically validate** existing hypotheses



Silent Voices

- 1350-1400
- +25 manuscripts
- produced at the monastery of Herne
- 13 scribes, working closely together on
 - translations from Latin to Dutch
 - copying texts for their own use
 - ... for other communities
 - ... for wealthy citizens of Brussels
- Marginal conversations!

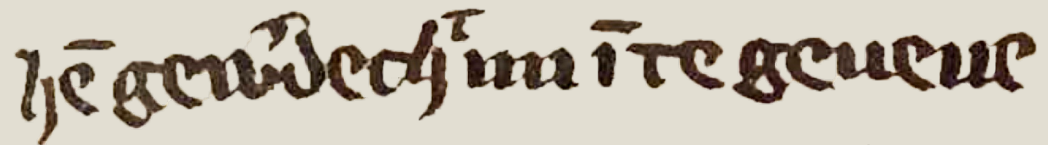




What are the copying strategies that made this group of monks such an incredibly well-oiled machine?



Editing the medieval text



hē gew'dech^t mi ī te geuene

Diplomatic transcription

hē gew'dech^t mi ī te geuene

Semi-Diplomatic transcription

hem gew^{er}dech^{eit} mi in te geuene

Normalised transcription

hem gewerdecheit mi in te gevene

Un-editing the medieval text

Edition	UnEdition
Printed on paper	Digitally in the cloud
Static — for eternity	Dynamic — until the next bugfix
Critical	(Hyper-)Diplomatic
Hide the sources	Foreground the sources
Old philology	New philology
Accuracy	Accessibility
Commercial	Open access

CD-ROM Middelnederlands



Wealth of Middle Dutch texts!

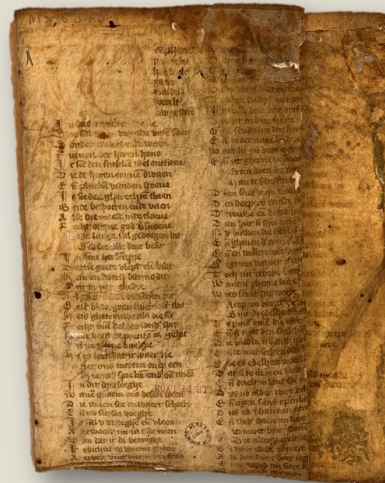
19th and 20th-century normalized editions...

Not all textual witnesses...

XML-TEI transcriptions

MVN framework by Herman Brinkman and Peter Boot (Huygens Institute)

- Options to keep or exclude the **markup** for capital letters, abbreviations, punctuation, etc.
- Lemmas, material aspects, ...
- **Advanced**
- **Legacy**



```
xml_1.xml x
MVN
13 <encodingDesc>
14 <xsi:include href="http://www.huygensinstituut.knaw.nl/project
15 <xmins:xsi="http://www.w3.org/2001/XMLSchema">
16 <xsi:fallback>
17 <xsi:include href="charDecl/charDecl.xml"/>
18 </xsi:fallback>
19 </xsi:include>
20 <editorialDecl>
21 <p><!-- Verantwoording van de editie --></p>
22 </editorialDecl>
23 </encodingDesc>
24 <profileDesc>
25 <handNotes>
26 <handNote xmlid="hi"><p>...</p></handNote>
27 </handNotes>
28 </profileDesc>
29 </teiHeader>
30 <text xmlid="Martijn">
31 <group><pb xmlid="L.fira" n="1ra"/>
32
33
34 <text n="Eerste Martijn" xmlid="M1">
35 <body>
36 <p><lb n="1" xmlid="L.fira.1"/><l n="001"><hi rend="capitalsize">W</hi>
```

Jacob van Maerlant, Martijn

Teksten: [Eerste Martijn](#) [Tweede Martijn](#) [Derde Martijn](#)

Foliumzijden: [1ra](#) [1rb](#) [1va](#) [1vb](#) [2ra](#) [2rb](#) [2va](#) [2vb](#) [3ra](#) [3rb](#) [3va](#) [3vb](#) [4ra](#) [4rb](#) [4va](#) [4vb](#)

ms>
Tekst Eerste Martijn

1	W	Aphene merten hoe saelt gaen	001
2	Sal dese werelt lange staen		002
3	In dus crancken loeue		003
4	Soe sal mijn vrouwe veren saen		004
5	Sonder twiuel ende waen		005
6	Rumen der heren houe		006
7	Ic sie den scalcken wel ontfaen		007
8	Die de heren coninc dwaen		008
9	Ende plucken vanden stoeue		009
10	Ic sie den gherechten slaen		010
11	Beide bespotten ende vaen		011
12	Alse die meese inde cloeue		012
13	Recht oftene god verscroeue		013
14	HOe lange sal gedoen dit		014
15	God die alle dinc besit		015
16	In sine herscapie		016
17	Dattie goede vleyt ende bidt		017
18	Hem en doech dat no dit		018



Character encodings with MUFI

Medieval Unicode Font Initiative

MUFI characters that are part of the Private Use Area (PUA), cause for problems...



Ñ	1E44	LATIN CAPITAL LETTER N WITH DOT ABOVE
ñ	1E45	LATIN SMALL LETTER N WITH DOT ABOVE
Ṅ	1E46	LATIN CAPITAL LETTER N WITH DOT BELOW
ṅ	1E47	LATIN SMALL LETTER N WITH DOT BELOW
ņ	A774	LATIN SMALL LETTER NUM
N̄	E1DC	LATIN CAPITAL LETTER N WITH HIGH MACRON (ABOVE CHARACTER)
ñ̂	E5D7	LATIN SMALL LETTER N WITH CIRCUMFLEX
ñ̄	E5DC	LATIN SMALL LETTER N WITH MEDIUM-HIGH MACRON (ABOVE CHARACTER)
ņ̇	E5EE	LATIN SMALL LETTER N WITH RING BELOW

Latin Small Letter N With Medium-High Macron
(uE5DC): wrong visualisation

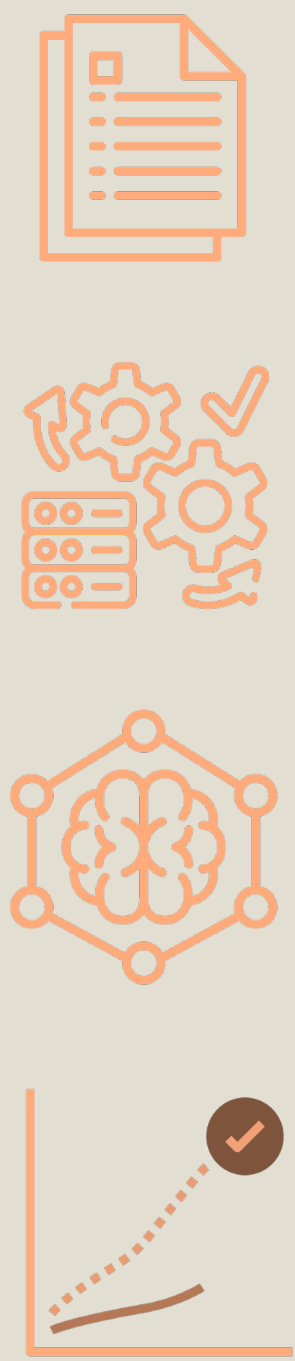
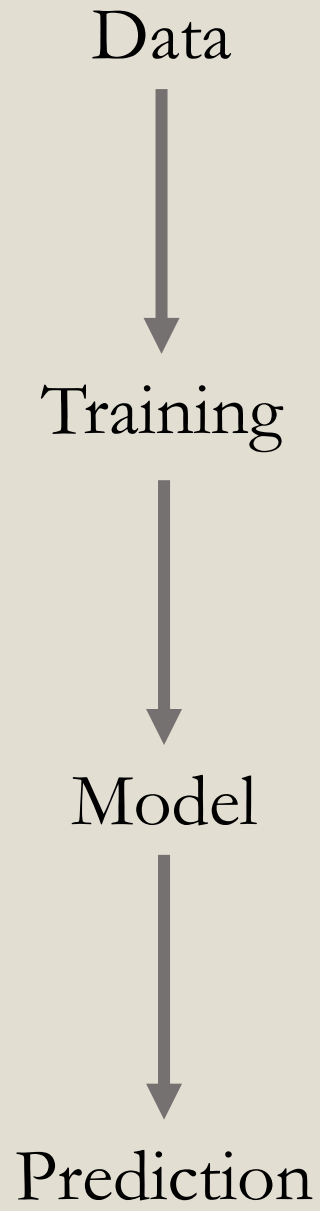
A	dat	neen	wart	ja	e	□	ia	-	neen
B	dat	neen	was	ja	e	□	ja	-	neen
C	dat	neen	was	ja	e	□	ja	was	neen
D	ende	neen	was	ia	ende	-	ia	was	neē
F	dat	neen	w't	ia	e	□	ia	-	neen
G	d3	neen	was	ia	e	□	ia	was	neē
O	dat	neen	was	ja	e	□	ia	-	neen
Y	dat	neen	was	-	-	-	ia	-	neen

Combining macron (u0304): wrong collation

A	dat	neen	wart	ja	en	-	ia	-	neen	-
B	dat	neen	was	ja	en	-	ja	-	neen	-
C	dat	neen	was	ja	en	-	ja	was	neen	-
D	ende	neen	was	ia	ende	-	ia	was	nee	-
F	dat	neen	w't	ia	en	-	ia	-	neen	-
G	d3	neen	was	ia	en	-	ia	was	nee	-
O	dat	neen	was	ja	en	-	ia	-	neen	-
Y	dat	neen	was	-	-	-	ia	-	neen	-

Latin Small Letter N With Tilde (u00F1): pragmatic
solution

A	dat	neen	wart	ja	eñ	ia	-	neen
B	dat	neen	was	ja	eñ	ja	-	neen
C	dat	neen	was	ja	eñ	ja	was	neen
D	ende	neen	was	ia	ende	ia	was	neē
F	dat	neen	w't	ia	eñ	ia	-	neen
G	d3	neen	was	ia	eñ	ia	was	neē
O	dat	neen	was	ja	eñ	ia	-	neen
Y	dat	neen	was	-	-	ia	-	neen



Transkribus Expert Client v1.25.0.7-SNAPSHOT (07_02_2023_16:38), Loaded doc: TRAINING_VALIDATION_SET_BigMiddleDutchModel_v2, ID: 133306

Search current document... 7 /133 Ground Truth

The screenshot shows a manuscript page with text in Gothic script. The text is primarily black, with a large initial 'S' and some words in red. There are numbered annotations (1-12) in blue and red circles. The text is arranged in two columns. The first column contains the main text, and the second column contains smaller text, possibly a commentary or a different version of the text.

1-2 bē·en·selē·xlviij ↩

1-3 Sonderlighe ↩

1-4 salmē·dese·quaetheit·altemale ↩

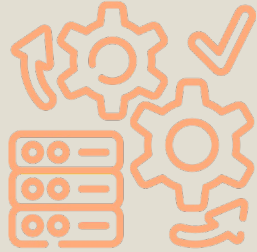
1-5 te·nieute·doen·vā·dē·cloest'e··dat·niemā·hē·en ↩

3,000 pages



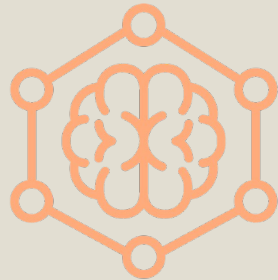
Risk: are these 3,000 pages a representative sample of the entire corpus?

Training



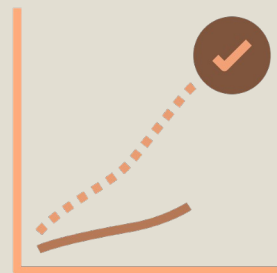
Risk management: make sure training material is well-stratified across:

Model



- all manuscripts
- different handwriting styles
- topics
- text layouts
- image resolutions
- etc.

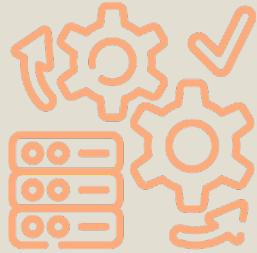
12,000 pages



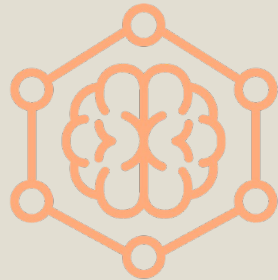
3,000 pages



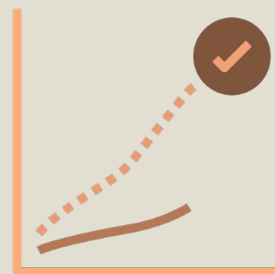
Training



Model



12,000 pages



	Character Error Rate (CER)	
	Train	Test
BigMiddleDutchModel	2.00%	2.70%

3,000 pages



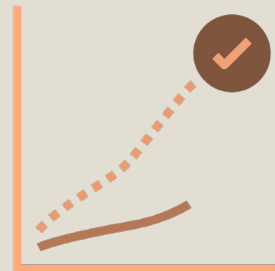
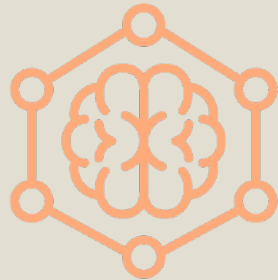
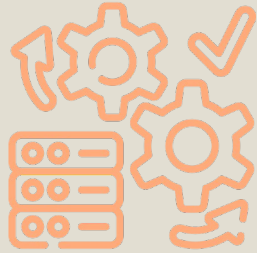
Training



Model



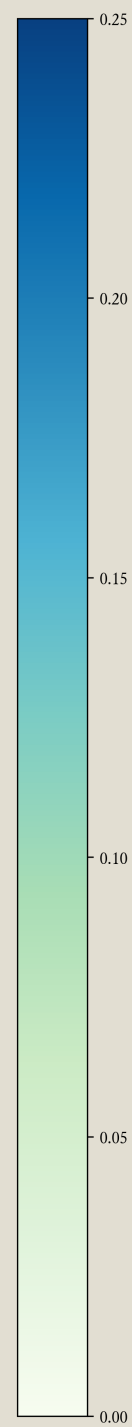
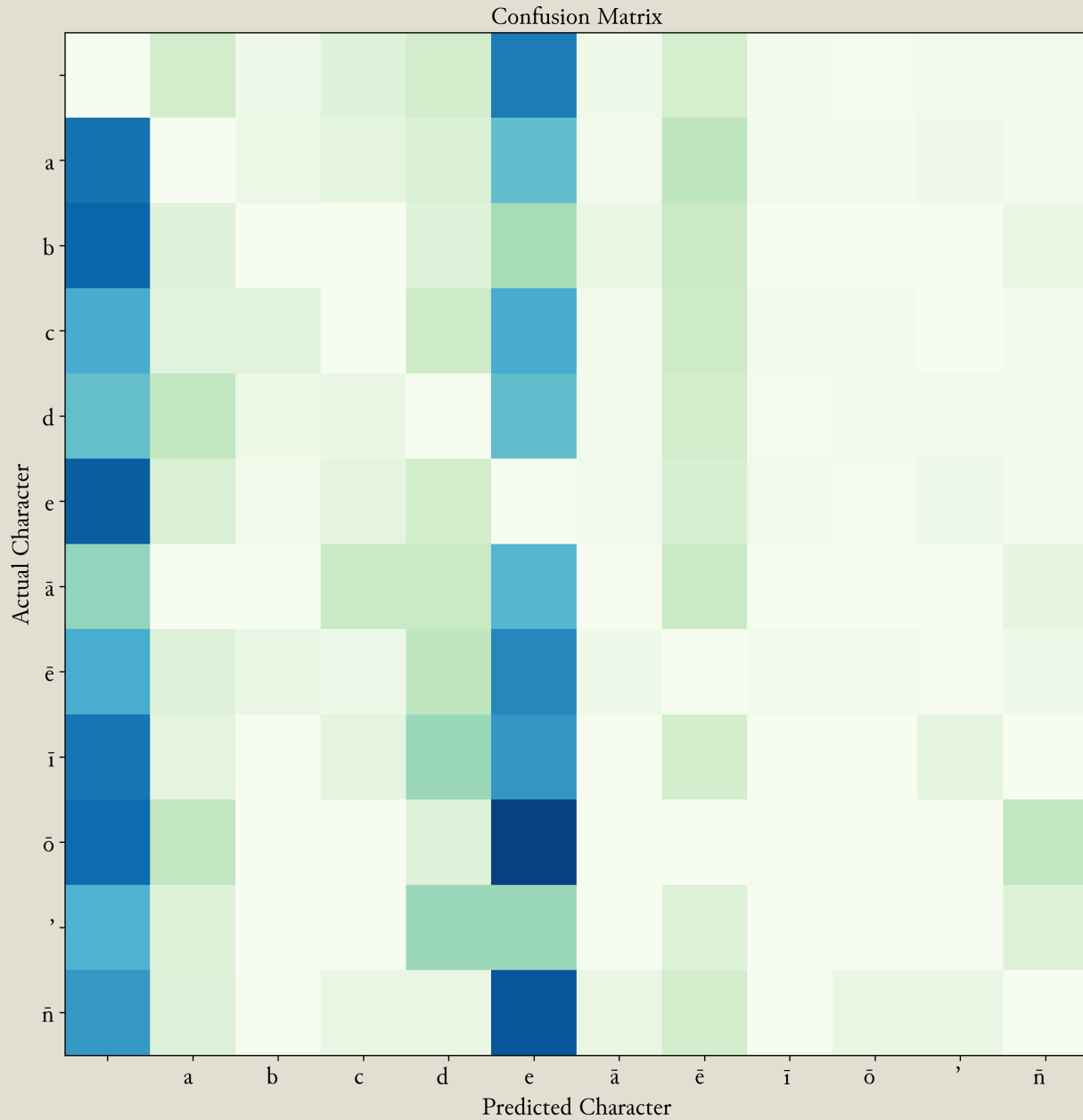
12,000 pages



	Character Error Rate (CER)	
	Train	Test
BigMiddleDutchModel	2.00%	2.70%

Is this good? Probably.
But not very informative...





Character Error Rate (CER)		
	Train	Test
BigMiddleDutchModel	2.00%	2.70%

References

- **Dekker, R. Haentjens, van Hulle, D., Middell, G., Neyt, V., & van Zundert, J.** (2015). Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project. *Digital Scholarship in the Humanities*, 30(3), 452–470. <https://doi.org/10.1093/llc/fqu007>
- **Driscoll, M. J.** (2010). The words on the page: Thoughts on philology, old and new. In J. Quinn & E. Lethbridge (Red.), *Creating the medieval saga: Versions, variability, and editorial interpretations of Old Norse saga literature*. (pp. 85–102). Syddansk Universitetsforlag.
- **Gueville, E., & Wrisley, D. J.** (z.d.). *Transcribing Medieval Manuscripts for Machine Learning*.
- **Haverals, W., & Kestemont, M.** (z.d.). From exemplar to copy: The scribal appropriation of a Hadewijch manuscript computationally explored. 20.
- **Morreale, L. & Albritton B.** (2021), Community, Collaboration, and the UnEdition. URL: <https://youtu.be/rpL2KY8Bk8E>
- **Kestemont, M.** (2018). Aan de taal kent men de hand. *Spiegel der Letteren*, 3, 157–188. <https://doi.org/10.2143/SDL.60.3.3285821>
- **Pierazzo, E.** (2016). Textual Scholarship and Text Encoding. In S. Schreibman, R. G. Siemens, & J. M. Unsworth, *A New Companion to Digital Humanities* (pp. 307–321). Wiley-Blackwell.
- **Rubin, D. C.** (1995). *Memory in oral traditions: The cognitive psychology of epic, ballads, and counting-out rhymes*. Oxford University Press.
- **Thaisen, J.** (2014). Initial position in the Middle English verse line. *English Studies*, 95(5), 500–513.
- **Van Dalen-Oskam, K.** (2012). The secret life of scribes. Exploring fifteen manuscripts of Jacob van Maerlant's *Scolastica* (1271). *Literary and Linguistic Computing*, 27(4), 355–372. <https://doi.org/10.1093/llc/fqs034>

racking the code

Challenges of Encoding Medieval Manuscripts

 Sofie  Moors (FWO PhD, University of Antwerp)

 Wouter  Haverals (FWO Postdoc, University of Antwerp)

Get in touch!

- Sofie.Moors@uantwerpen.be
- <https://github.com/SofieMoors>
- Wouter.Haverals@uantwerpen.be
- <https://github.com/WHaverals/>