racking the code

Challenges of Encoding Medieval Manuscripts

Sofie Moors (FWO PhD, University of Antwerp)

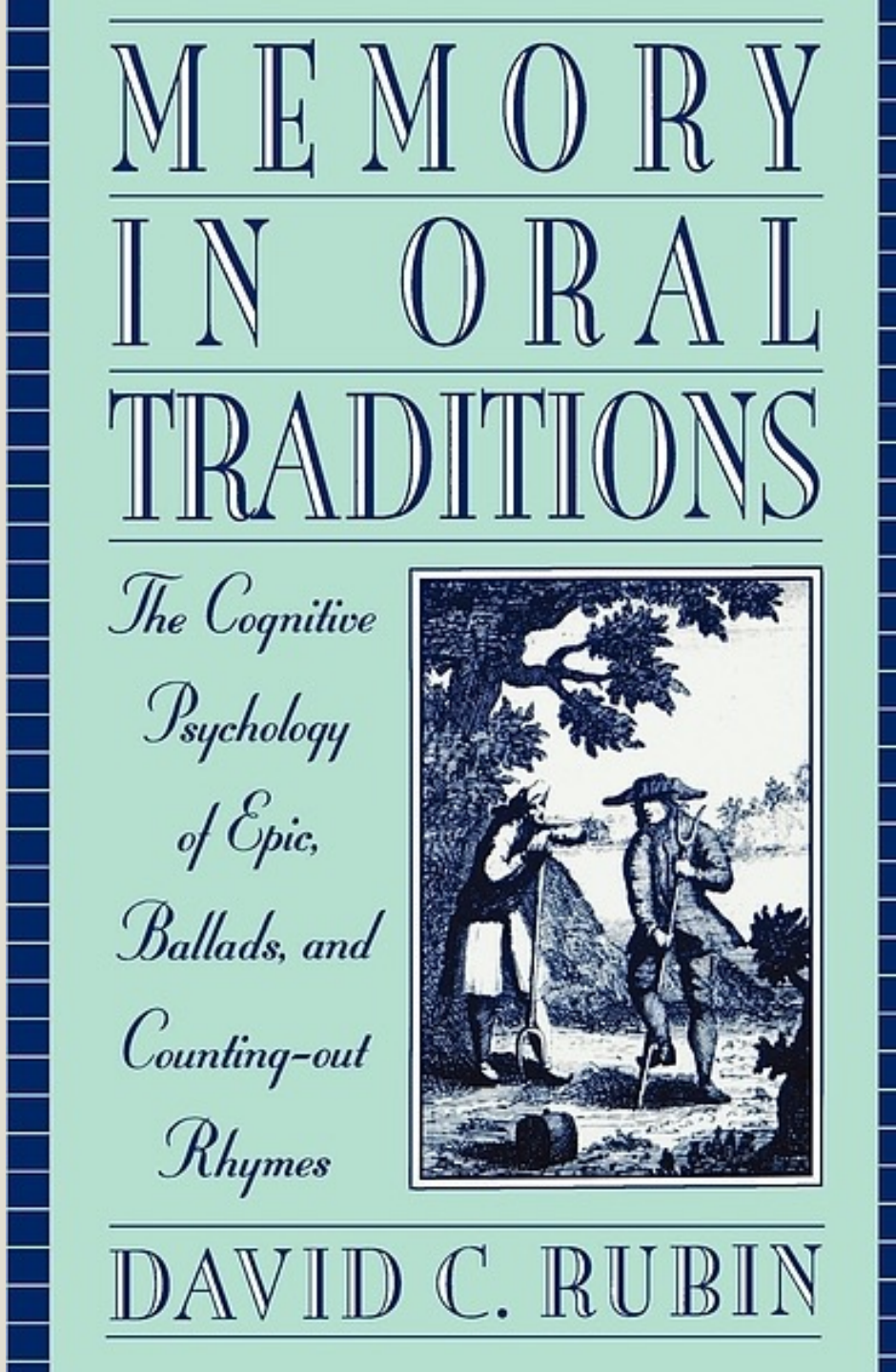
Wouter Haverals (FWO Postdoc, University of Antwerp)

Variation is what the medieval text is about. (Driscoll 2010)

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)



- Computational investigation of the transmission of ***Martijn Trilogy*** by Jacob van Maerlant

- **Computational alignment** of 17 different manuscripts, fragments, and prints before 1500

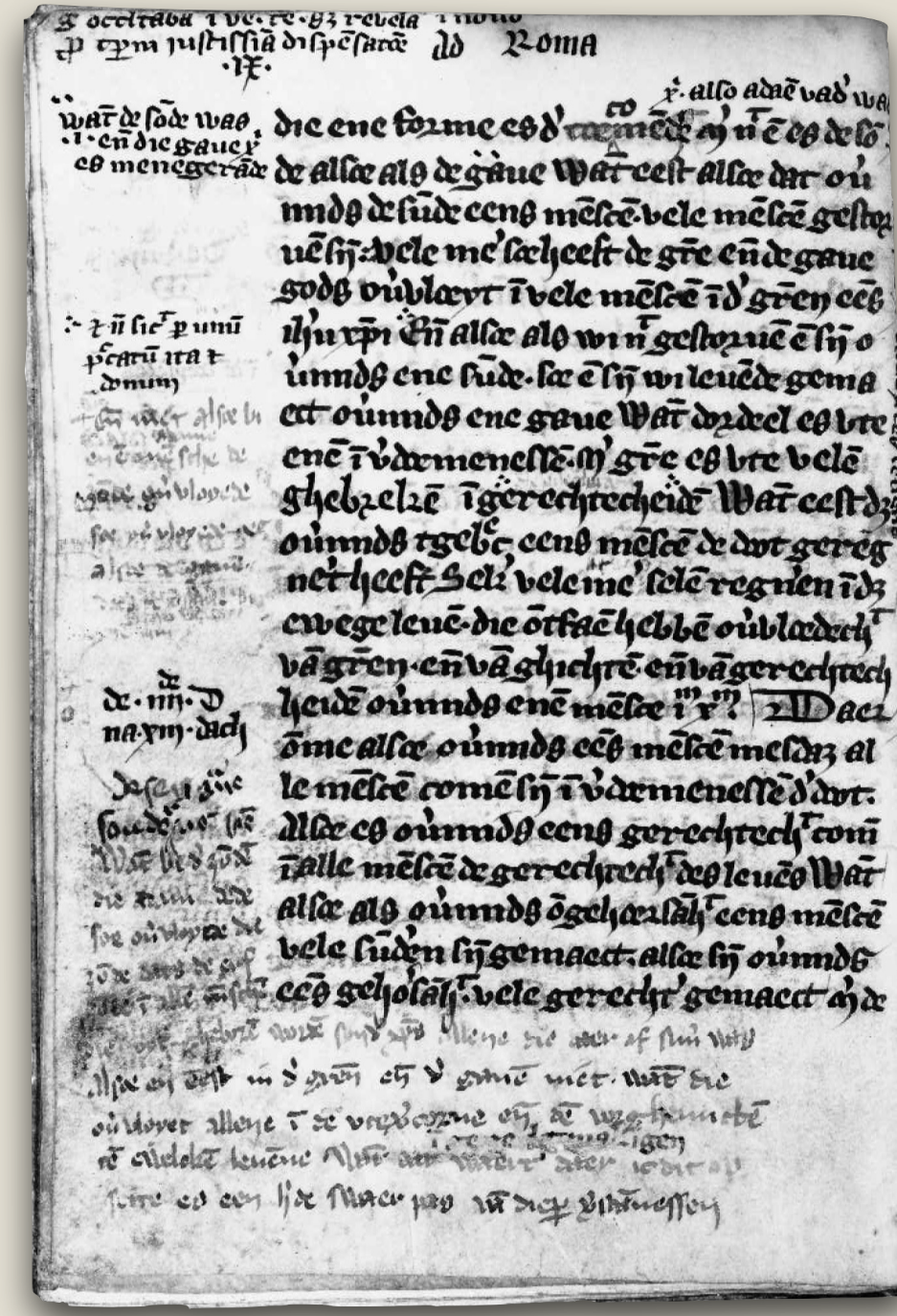
- **Empirically validate** existing hypotheses

Silent Voices

- 1350-1400
- +25 manuscripts
- produced at the monastery of Herne

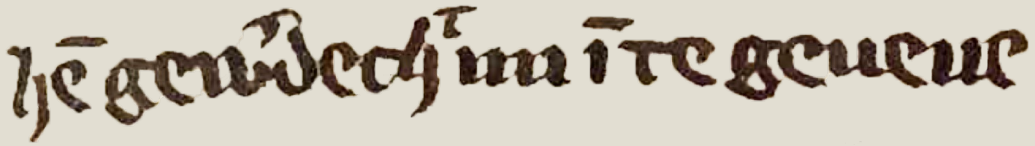
- 13 scribes, working closely together on
  - translations from Latin to Dutch
  - copying texts for their own use
  - … for other communities
  - … for weatlhy citizens of Brussels

- Marginal conversations!



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

Editing the medieval text



Diplomatic transcription
hē gew’dechᵗ mi ī te geuene

Semi-Diplomatic transcription
he*m* gew*er*dech*eit* mi i*n* 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!

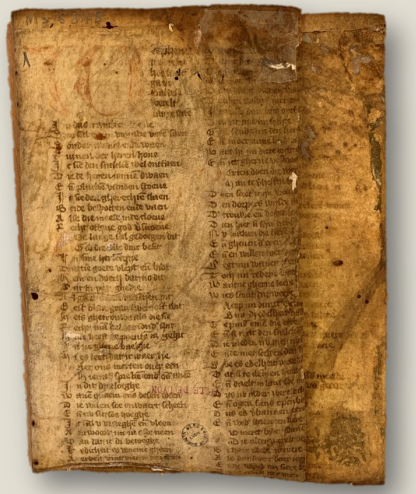
19th and 20th-century normalized editions…

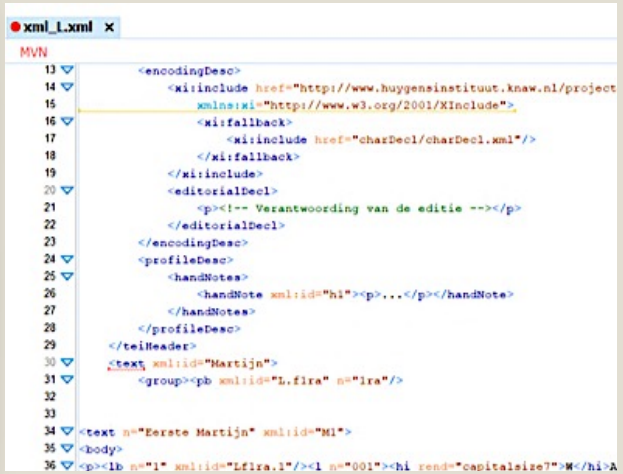
Not all textual witnesses…

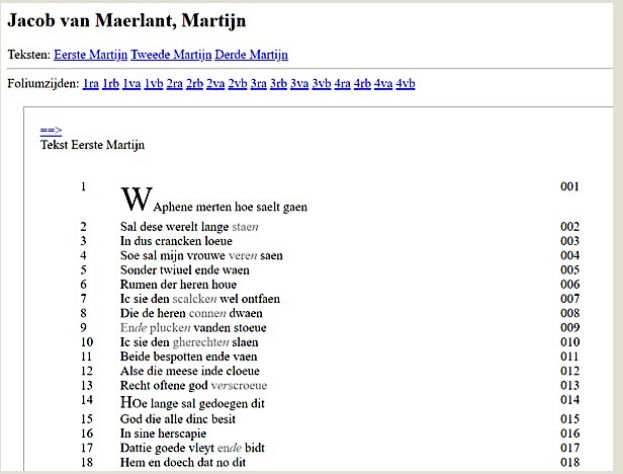
XML-TEI transcriptions

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







Character encodings with MUFI

Medieval Unicode Font Initiative

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

Ṅ	**1E44**	LATIN CAPITAL LETTER N WITH DOT ABOVE
ṅ	**1E45**	LATIN SMALL LETTER N WITH DOT ABOVE
Ṇ	**1E46**	LATIN CAPITAL LETTER N WITH DOT BELOW
ṇ	**1E47**	LATIN SMALL LETTER N WITH DOT BELOW
n⊣	**A774**	LATIN SMALL LETTER NUM
N̄	**E1DC**	LATIN CAPITAL LETTER N WITH HIGH MACRON (ABOVE CHARACTER)
n̂	**E5D7**	LATIN SMALL LETTER N WITH CIRCUMFLEX
n̄	**E5DC**	LATIN SMALL LETTER N WITH MEDIUM-HIGH MACRON (ABOVE CHARACTER)
n̥	**E5EE**	LATIN SMALL LETTER N WITH RING BELOW

Alignment using **CollateX software**

(Dekker and Middell 2011)

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	dʒ	neen	was	ia	e	□	ia	was	neē
O	dat	neen	was	ja	e	□	ia	–	neen
Y	dat	neen	was	–	–	–	ia	–	neen

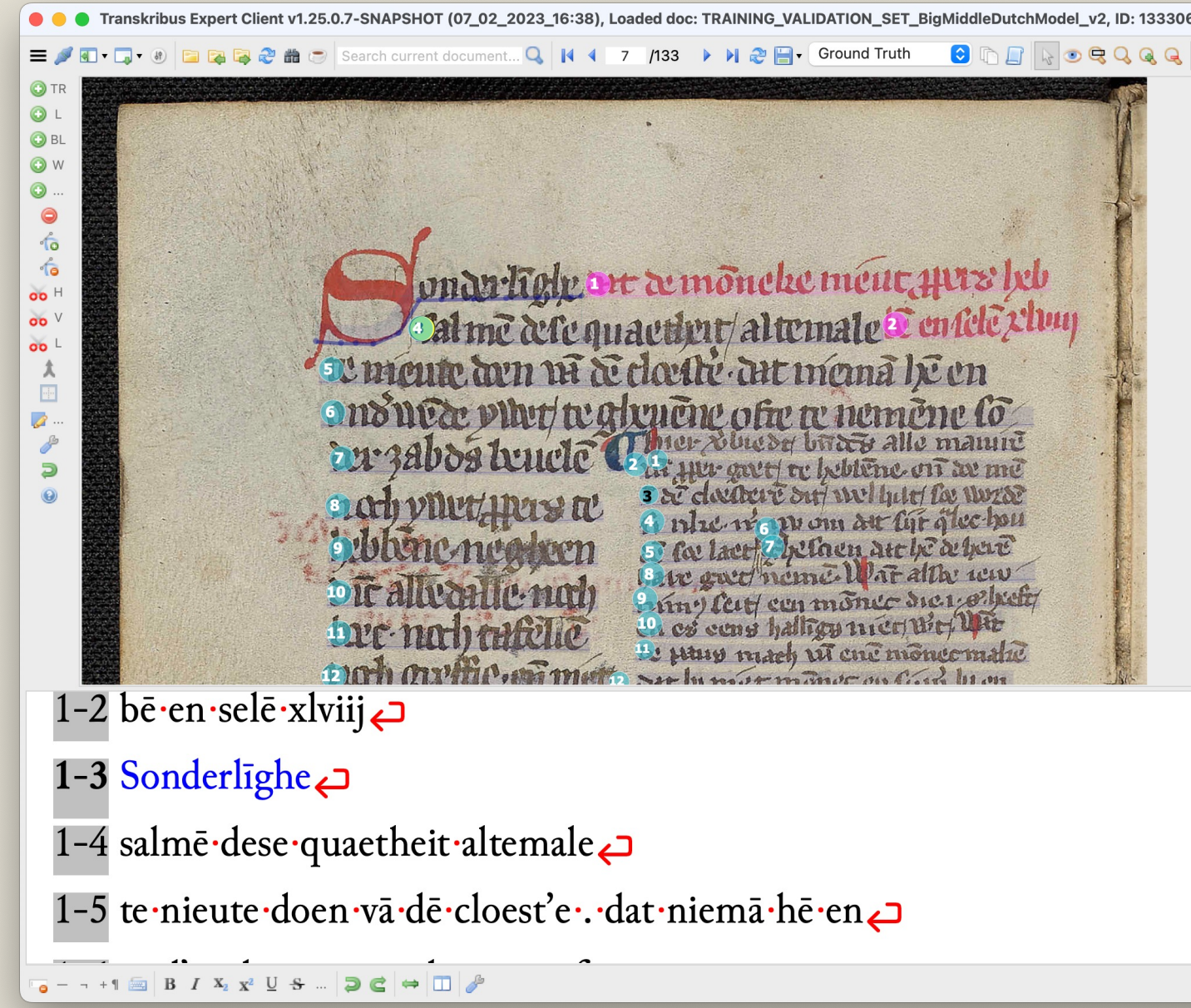
Latin Small Letter N With Tilde (u00F1): pragmatic solution 

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	dʒ	neen	was	ia	eñ	ia	was	neē
O	dat	neen	was	ja	eñ	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	dʒ	neen	was	ia	en	-̄	ia	was	nee	-̄
O	dat	neen	was	ja	en	-̄	ia	–	neen	–
Y	dat	neen	was	–	–	–	ia	–	neen	–

- In the Silent Voices project:
  - Ca. 15,000 manuscript folios to be transcribed
  - Relying on the popular Handwritten Text Recognition tool *Transkribus*

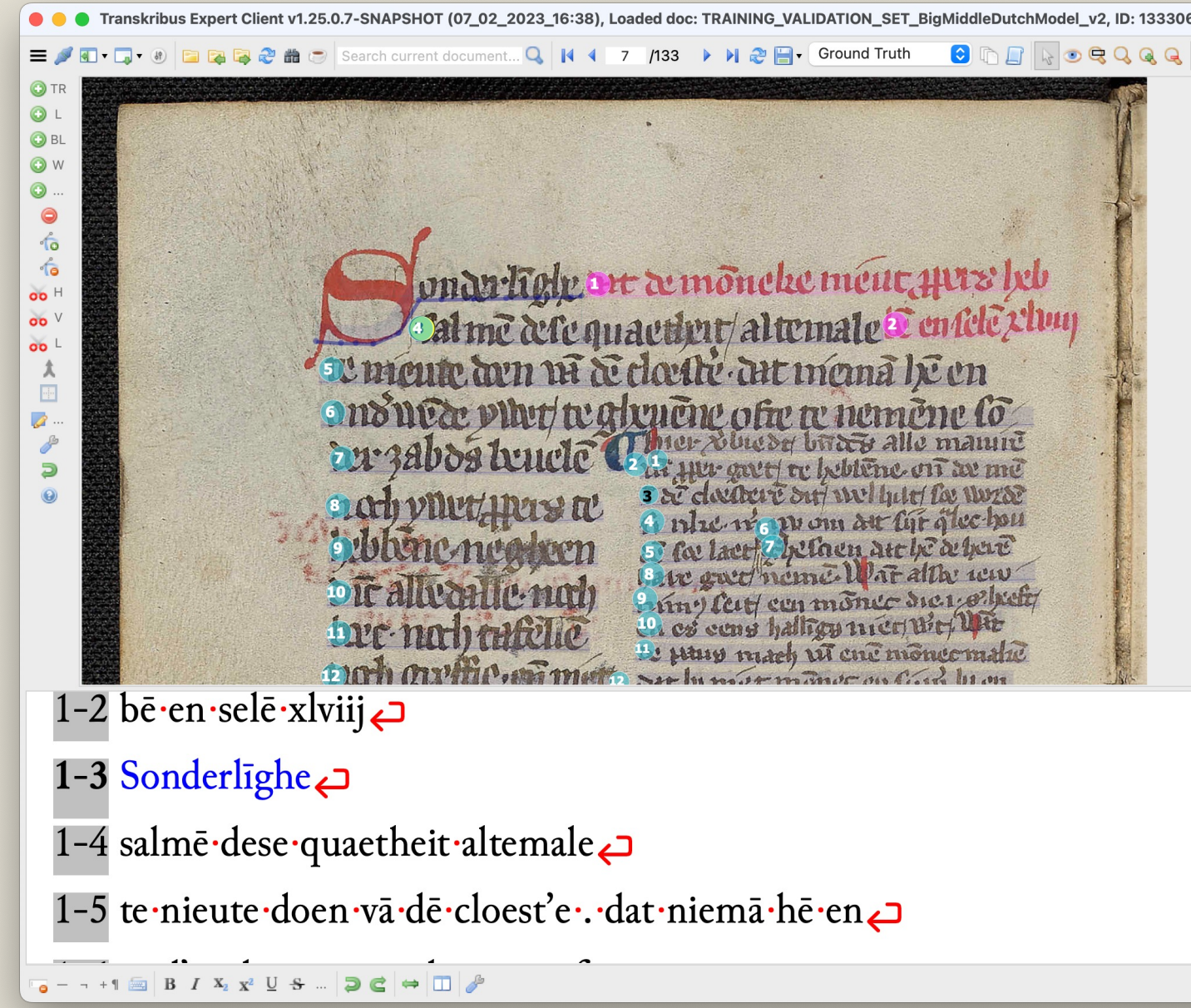


Data

Training

Model

Prediction

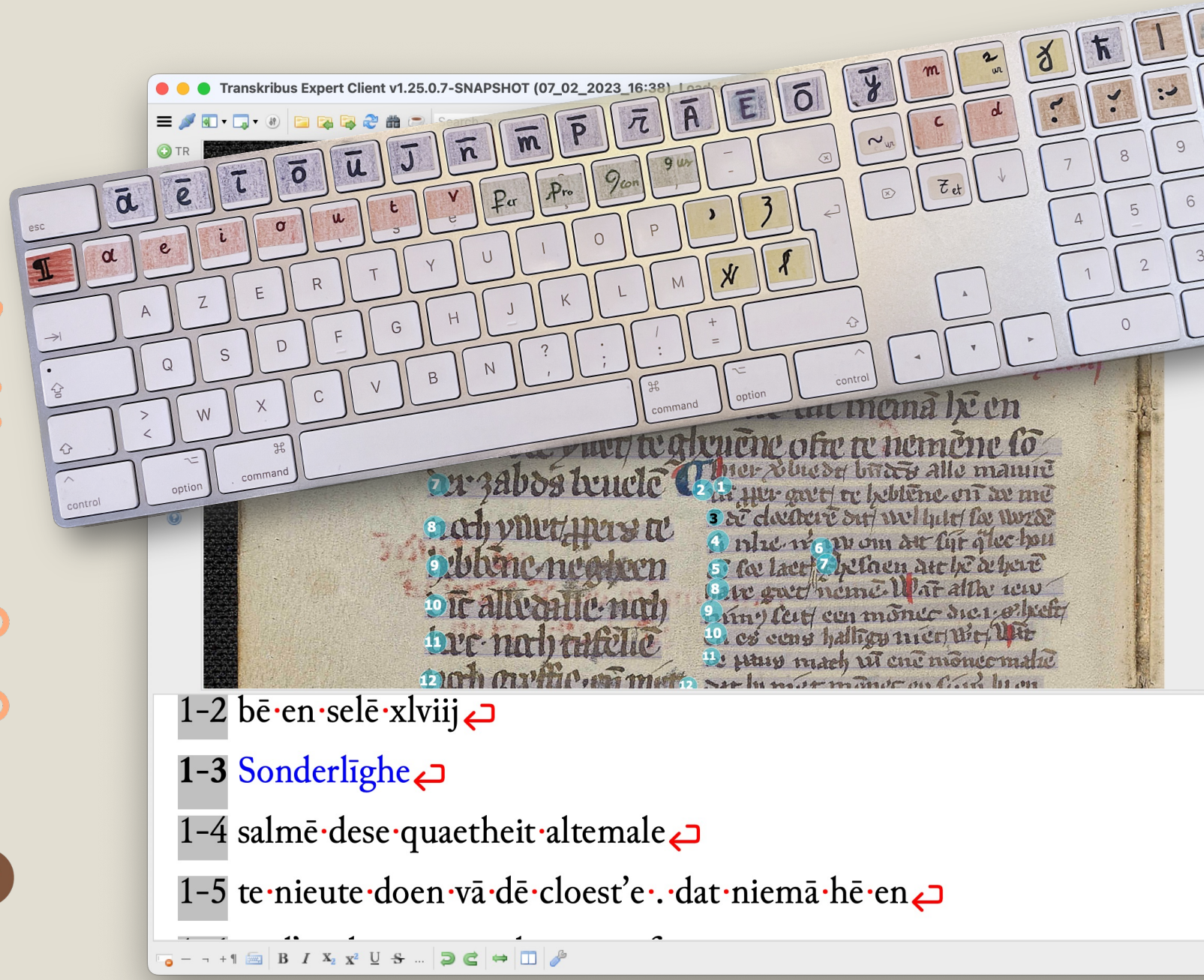


3,000 pages

Training

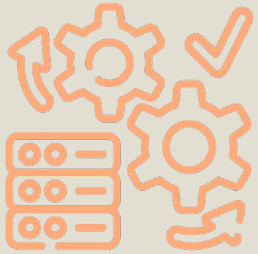
Model

12,000 pages

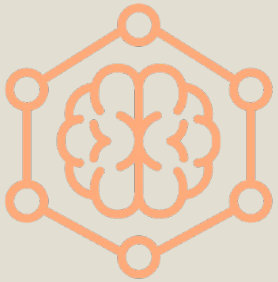


3,000 pages

Training



Model



12,000 pages

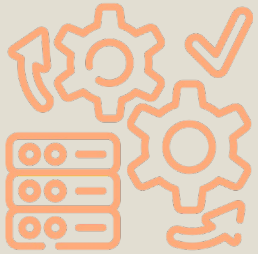
Risk: are these 3,000 pages a representative sample of the entire corpus?

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

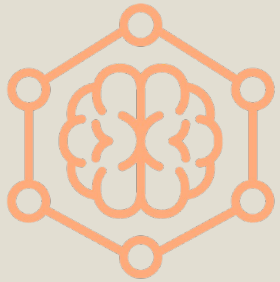
- all manuscripts
- different handwriting styles
- topics
- text layouts
- image resolutions
- etc.

3,000 pages

Training



Model

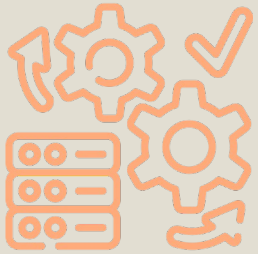


12,000 pages

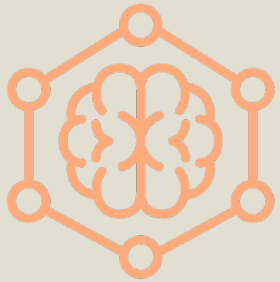
	Character Error Rate (CER)	
	Train	Test
BigMiddleDutchModel	2.00%	2.70%

3,000 pages

Training



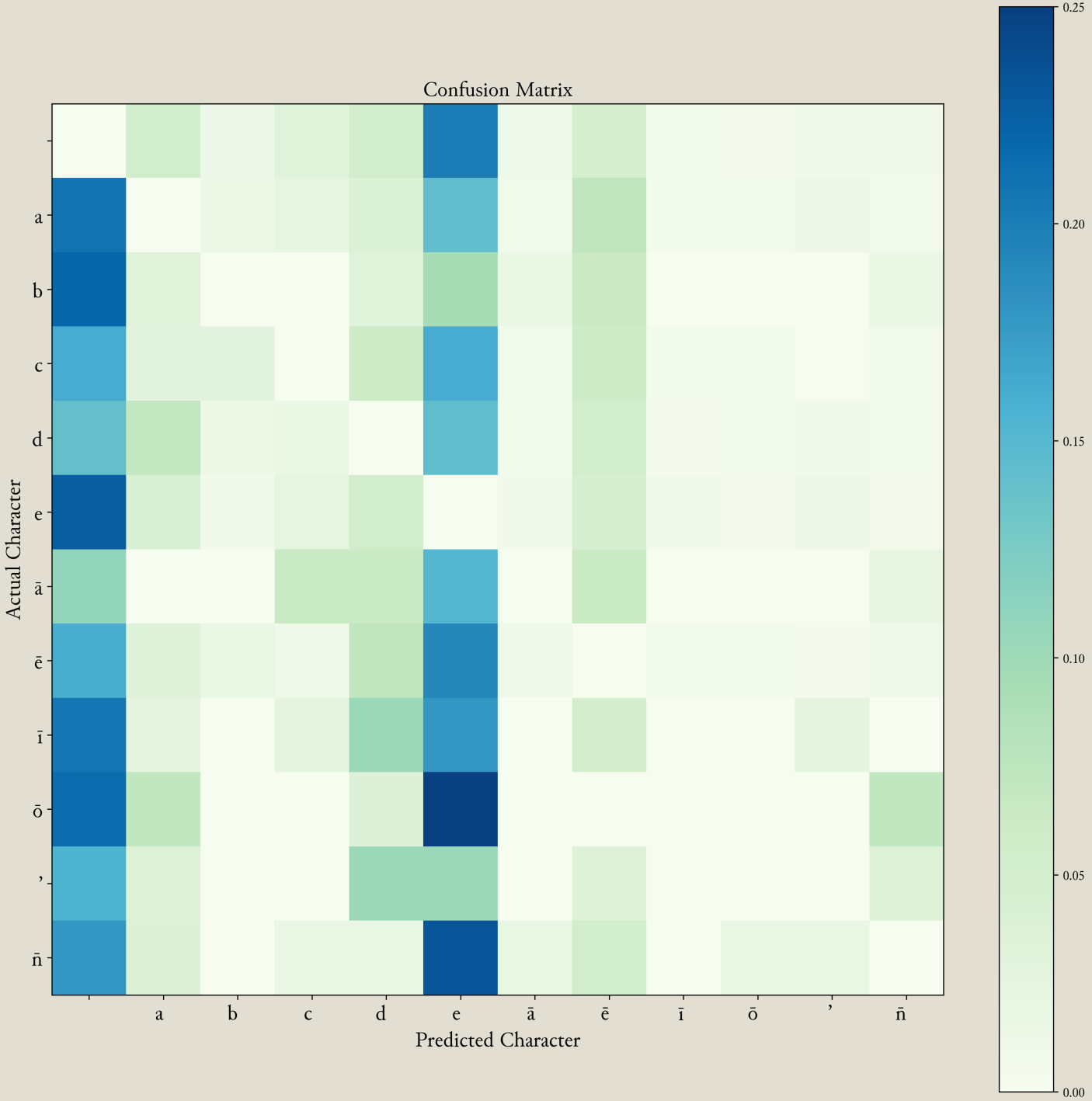
Model



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

Cracking 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/