The Middle Dutch Manuscripts Surviving from the Carthusian Monastery of Herne (14th century): Constructing an Open Dataset of Digital Transcriptions

Wouter Haverals[1,2,3], Mike Kestemont[2,3,*]

[1]*Center for Digital Humanities (CDH), Princeton University, USA*

[2]*Antwerp Center for Digital Humanities and Literary Critcism (ACDC), University of Antwerp, Belgium*

[3]*Institute for the Study of Literature in the Low Countries (ISLN), University of Antwerp, Belgium*

Abstract

A substantial collection of Middle Dutch manuscripts survives from the Carthusian monastery of Herne (*Hérinnes-lez-Enghien*) in nowadays Belgium. During the latter half of the fourteenth century, Herne served as a significant literary hotspot in the region around Brussels, with a devoted community of monks deeply involved in the production of (vernacular) texts and manuscripts, often as collaborative efforts. The corpus offers abundant material for the (computational) exploration of authorship, translation, and scribal cultures in the premodern Low Countries. Yet, much of this material has remained digitally inaccessible. Here we describe the creation of an almost exhaustive, open-access dataset comprising diplomatic transcriptions of all known Middle Dutch Herne manuscripts, acquired through handwritten text recognition. Apart from rich codicological and textual metadata, we include a normalized text layer (with expanded abbreviations), as well as a linguistic annotation layer (with lemmas and part of speech tags). We conclude by discussing our work against current trends in medievalist scholarship. The dataset is released together with this paper and we encourage its re-use in future research.

Keywords

Carthusians, medieval literature, handwritten text recognition, scribal cultures, orthography

1. Herne and Middle Dutch literature

A sizable corpus of Middle Dutch literature survives today, composed in the various vernacular Germanic dialects that were spoken in the Low Countries during the middle ages (ca. 1200-1450). These works were manually copied onto handwritten text witnesses, mostly manuscripts or book rolls, which are nowadays often only extant in a fragmentary or heavily damaged state. In many cases, very little is known concerning the original provenance of such books, including details about the scribes, patrons, and the date and place of composition. Especially for vernacular manuscripts, it is challenging to connect the remaining pieces of evidence.

Against this backdrop, the Herne corpus stands out: an extensive collection of Middle Dutch manuscripts that were all produced in a single environment, by a restricted number of intensively collaborating scribes, during the limited time period of the second half of the fourteenth century.

Founded in 1314 as the first Carthusian monastery in the Low Countries, the monastery in Herne (Fr. *Hérinnes-lez-Enghien*) is located roughly thirty kilometers southwest of Brussels. The monastery actively liaised with cultural stakeholders in the city; as indicated by some prologues, local citizens acted as patrons for specific projects, while the monks engaged in commercial (*pro pretio*) book production, delivering, or correcting copies for Brussels book dealers, for instance. This heightened level of book production at Herne should be understood against the Carthusian monastic rules that promoted silence [15]. The monks spent most of their time in individual cells and were restricted from verbal communication except during specific moments each week. As these rules of silence and minimal outside interaction precluded Carthusians from traditional preaching, the (silent) production and dissemination of written texts became their principal form of apostolate.

The Carthusian orientation of the Herne monastery influenced their text and book production in specific ways [26]. Firstly, because the monks worked in isolation, Carthusian libraries are characterized by the presence of multiple copies of important texts (so that they could consult these works simultaneously). Secondly, because the monks were silent, collaborations were often discussed in writing: many manuscripts contain marginal notes in which collaborating scribes discussed the quality of a text or negotiated translations. Such detailed paratexts are extremely rare for this period, providing an exceptional lens to study medieval scribal cultures. Thirdly, while the (still young) monastery also produced many books for itself, an important share of the manuscripts were in fact intended for lay readers beyond its walls. This helps explain the fourth defining characteristic of the Herne corpus, namely its focus on the vernacular. The monastery acted as a hub for the translation and distribution of Middle Dutch literature, part of its apostolate involved making Latin texts accessible to a lay audience. A remarkable achievement, for instance, was the Herne Bible translation initiated around 1360 [24, 8]. This project became the first near-exhaustive vernacular Bible translation in prose in history, a venture both novel and controversial in the fourteenth century. Many of the surviving prologues discuss the backlash faced by the translators.

While mostly religious texts from Herne survive, it is clear that the Carthusians had access to a broad spectrum of contemporary material, including a variety of worldly texts [14]. For literary history, Herne holds special importance as it is home to many rare works, including the two earliest manuscripts of the world-renowned mystical poetess Hadewijch. Moreover, many *unica* are only extant in Herne copies, such as strophic poems by Jacob van Maerlant or the second part of the Middle Dutch adaptation of the *Speculum historiale* [11]. Finally, Herne's history is intertwined with the biography of John of Ruusbroec, another famous mystical author (and, after Anne Frank, the most often translated Dutch author) [39, 32]. The Carthusians had gained access to an unauthorized copy of an early work of Ruusbroec and asked him for clarifications regarding some controversial passages. In 1362, despite his advanced age, Ruusbroec journeyed on foot to Herne to discuss and clarify controversial passages with the monks. His clarifications would eventually yield an important commentary (*Dat boecsken der verclaringhe*).

2. The surviving corpus of manuscripts

The attribution of the Herne corpus to this monastery only came about recently. Prior to 2002, the lion's share of the manuscripts discussed here were widely believed to have been produced in Rooklooster (Fr. *Rouge Cloître*), an Augustinian priory located on the eastern outskirts of Brussels and established in 1367. In his pioneering PhD, Erik Kwakkel was able to show that most of these books were, in fact, not copied in Rooklooster but in Herne [26]. He convincingly argued that only later, a surprisingly large share of the Herne books ended up in the possession of the Rooklooster monastery, which explains why so many of these manuscripts bare contemporary ownership marks pointing to the priory. Interestingly, these books arrived in Rooklooster via rather diverse routes and didn't travel in bulk [25]. Many of these specimens feature in a well-known book list (entitled *Die Dietsche boeke die ons toebehoeren* or "The Dutch books that belong to us"), although it is still unclear whether this inventory has been drawn up in Rooklooster or in Herne.

Kwakkel was able to attribute the production of a cohesive collection of manuscripts to the Herne charterhouse on the basis of a variety of arguments. Firstly, a number of scribal hands can be localized in Herne beyond reasonable doubts, because they have worked on documents that played a central role in the local community, such as the so-called Necrology Scribe (β), who contributed to the local necrologies (e.g. BRUSSELS, RL, 21536–40). Consequently, scribes who co-produced manuscripts with β can also be confidently located in Herne, such as the Speculum Scribe (α; see Fig. 1). Secondly, it seems that the Herne community employed a unique correction system, through placing a struck out delta in the margin. Given that this correction system appears exclusive to Herne during this period, the production of many other manuscripts can tentatively be attributed to the charterhouse.

All in all, Kwakkel has been able to identify 13 hands who contributed to Herne manuscripts in the second half of the fourteenth century. (These have been numbered 1 to 13, although e.g. scribe 1 and 2 are also frequently referred to as α and β.) These manuscripts, however, remain shrouded in anonymity: none of the local scribes are known by their personal names. Despite the existence of numerous highly informative prologues, the monks fiercely avoided self-identification in these texts and frequently referred to their work as a collective endeavor (e.g., through the use of the first-person plural pronoun). Only a small number of Carthusians lived in Herne at any given time; around 1390, for instance, we know the names of the 17 inhabitants who resided permanently in the community (both monks and lay brothers), including two brothers, Petrus and Jacob Naghel, who both took up authoritative positions in the community [26, p. 85]. Many scholars have attempted to identify Petrus Naghel as the 'Bible translator of 1360' [24], although this attribution has remained controversial – not in the least because many of the arguments that apply to Petrus equally apply to his brother Jacob [10]. Some observations can nevertheless be drawn from the scant evidence that is available: Petrus Naghel, for instance, died in 1395 and does cannot be identified as the hyper-productive Speculum Scribe (α), whose youngest (dated) booklet stems from 1402.

The research project "Silent voices: A Digital Study of the Herne Charterhouse as a Textual Community (ca. 1350-1400)", funded by the Flemish Research Agency (Belgium), studies the evolution of the scribal culture in the Herne monastery through computational text analysis [18]. For this project, limited to the vernacular book production in the charterhouse, we rely

Figure 1: Fol. 39v. from the Heber-Serrure codex (GHENT, UL, 1374) by the Speculum scribe (α). Note the very dense layout, with a rhyming text in in continuous notation and the high number of abbreviations. Available via IIIF through Ghent University Library [url]. Creative Commons Public Domain Mark 1.0.

Table 1

List of manuscripts included in the dataset. Datings based on the appendix in Kwakkel [26].

Codex signature	Local scribes	Size	Content	Date	Images
Brussels, RL, 1805-8	α, β, 3	68 fol.	Dialogues (Gregory the Great)	ca. 1395	Digital
Brussels, RL, 2485	β	69 fol.	Rule of Saint Benedict	ca. 1373	Digital
Brussels, RL, 2849-51	α, β	348 fol.	New Testament	1375–1400	Digital
Brussels, RL, 2877-2878	10	166 fol.	Hadewijch (MS B)	1350–1400	Digital
Brussels, RL, 2879-2880		101 fol.	Hadewijch (MS A)	1325–1350	Digital
Brussels, RL, 2905-9	α	101 fol.	Hore dochter	1375–1400	Digital
Brussels, RL, 2979	α	167 fol.	Gospels	ca. 1350	Digital
Brussels, RL, 3091	9	227 fol.	John of Ruusbroec	1350–1375	Digital
Brussels, RL, 3093-95	α, β	188 fol.	Der minnen gaert; Pseudo-Bernardus	1375–1400	Digital
Brussels, RL, 394-98	β, 5, 6 & 7	236 fol.	Rule of Saint Benedict	1373–1383	Catalogue
Ghent, UL, 1374	α	137 fol.	Spiegel historiael; Jacob van Maerlant	1375–1400	IIIF
Ghent, UL, 941		90 fol.	Hadewijch (MS C)	ca. 1350	IIIF
Paris, Bibl. de l'Arsenal, 8224		167 fol.	Heinrich Seuse's Horologium (transl.)	ca. 1350	Catalogue
Paris, BM, 920	α, 4	58 fol.	Ruusbroec; Hadewijch	1325–1400	Catalogue
Saint Petersburg, BAN, O 256	α	212 fol.	Gospels	1325–1350	
Vienna, ONB, Cod. 13.708	α	250 fol.	Spiegel historiael; Ruusbroec	1393–1402	IIIF
Vienna, ÖNB, SN 12.857	α, β, 9	230 fol.	Gospels	1375–1400	IIIF
Vienna, ÖNB, SN 12.905	9	210 fol.	Epistle pericopes	1350–1375	IIIF
Vienna, ÖNB, SN 65	α(?), 4	94 fol.	Heinrich Seuse's Horologium (transl.)	1375–1400	IIIF

on diplomatic transcriptions, obtained through handwritten text recognition. In this paper, we outline the construction of the underlying dataset, which is released in conjunction with this paper. The composition of the full corpus can be found in Table 1. The corpus is semi-exhaustive and encompasses almost all currently known Middle Dutch manuscripts associated with Herne (i.e. the manuscripts which were produced there or which were at least in the possession of the community). Non-vernacular manuscripts (e.g. the Latin codex Brussels, RL, 21536-40) are excluded, as well as very short texts (such as the aforementioned book list). Local scribes are identified through their Kwakkel number. Ghent, UL, 941 (Hadewijch MS C) and Paris, Bibl. de l'Arsenal, 8224 are additionally included: even though these were not produced in Herne, they stand in a close relationship to the corpus. Note that most of the manuscripts can only be dated approximately.

When considering the codices listed in Table 1, it is important to note that these are typically *composite* manuscripts, consisting of distinct codicological units, which were sometimes not even bound together until post-medieval times. Kwakkel adopts the term "production unit" in this context: "groups of quires that formed a material unity at the time of production. Such quires were copied "in one go", by either one or more scribes. A codex may contain several production units" [27, p. 13]. This distinction is crucial, because it is typically the production units that can be unambiguously dated. The larger codex Vienna, ÖNB, Cod. 13.708, for instance contains an older core that dates to 1393–1394, but also a younger extension that was only added in 1402 [26, p. 97–100]. The dataset presented here therefore contains detailed metadata on the production unit which each manuscript page belongs to.

3. Scribal characteristics

The contributions of the 13 local scribes vary in size and nature. Whereas the corpus contains a few, extremely large scribal oeuvres, many hands only appear sporadically. Table 2 shows for each manuscript witness, the number of production units (containing Middle Dutch), folium sides, number of glyphs that are available, alongside the number of unique scribal hands. In handwritten text recognition (HTR), the size and diversity of the annotated training material are known to have a meaningful impact on the model quality. We therefore anticipate lower HTR performance for less common scribal hands. Note that this Carthusian corpus contains several parallel transcriptions of the same works: these pose interesting cross-source challenges for HTR. On the one hand, transcribing a new version of a text that has been observed during training might increase the recognition accuracy. On the other hand, there's also a risk of the engine overfitting to the previous copy and ignoring variations present in the new one.

Table 2
Summary statistics for Middle Dutch manuscripts in the corpus.

Witness	Production units with Middle Dutch text	Folia recto + verso	Number of tokens	Type-token ratio	Number of unique chars.	Number of unique scribes
Brussels, RL, 1805-1808	3	132	95,503	6.73	107	4
Brussels, RL, 2485	2	135	27,442	5.99	87	1
Brussels, RL, 2849-51	7	693	154,399	8.10	122	1
Brussels, RL, 2877-78	2	330	90,181	7.61	84	1
Brussels, RL, 2879-80	3	202	82,113	7.43	93	3
Brussels, RL, 2905-09	10	384	46,954	5.73	94	2
Brussels, RL, 2979	2	49	216	1.61	45	3
Brussels, RL, 3091	1	452	75,179	7.69	96	2
Brussels, RL, 3093-95	2	374	40,444	5.19	101	2
Ghent, UL, 1374	5	264	77,717	7.28	116	1
Ghent, UL, 941	2	91	29,563	5.51	75	2
Paris, Bibl. Mazarine, 920	8	285	54,893	5.34	97	7
Paris, Bibl. de l'Arsenal, 8224	2	22	1020	1.66	61	3
Saint Petersburg, BAN, O 256	5	172	43,375	4.96	95	3
Vienna, ÖNB, 12.857	4	479	84,063	7.73	104	4
Vienna, ÖNB, 12.905	1	400	73,196	6.13	96	2
Vienna, ÖNB, 13.708	11	490	281,228	8.30	118	1
Vienna, ÖNB, 65	4	186	74,160	6.34	92	2

Another important factor is the intended readership of a codex which correlates with the quality of its execution. Local scribes were able to adapt the quality of a codex and their writing style as a function of the anticipated audience – for various scribes, Kwakkel distinguishes between a low, medium or high 'style' of execution (see Fig. 2 for examples). The codex Brussels, RL, 394–98, for instance, was intended for a (female) Benedictine sister in Vorst (Fr. *Forest*), and thus explicitly commissioned for the outside world. This attractive manuscript offers high readability, with a generous leaf size, spacious layout, and fewer abbreviations to enhance content accessibility, evidently to enhance the accessibility of its contents. Manuscripts which were likely meant to function in-house only adopt much more abbreviations. At the other end of the spectrum, we find the small-size miscellany Ghent, UL, 1374, created entirely from leftover low-grade parchment, which appears to have functioned as a personal booklet for the scribe. The abbreviation frequency is particularly high in this manuscript with its densely packed layout. Such aspects of readability and accessibility will inevitably influence the quality of the automatic transcriptions discussed below.
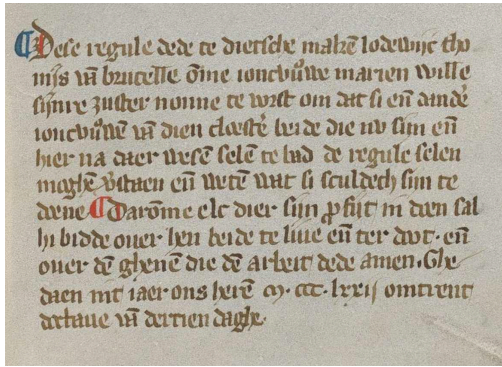
(a) High style: fol. 4r (cutout) in Brussels, RL, 2485, which was meant to function outside of Herne.
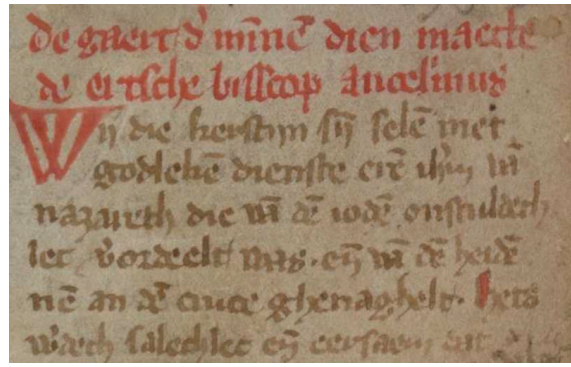
(b) Medium style: fol. 1r (cutout) in Brussels, RL, 3093-95, which likely functioned locally inside the charterhouse.

Figure 2: Examples of the 'medium' and 'high' execution styles of the Necrology scribe (β).
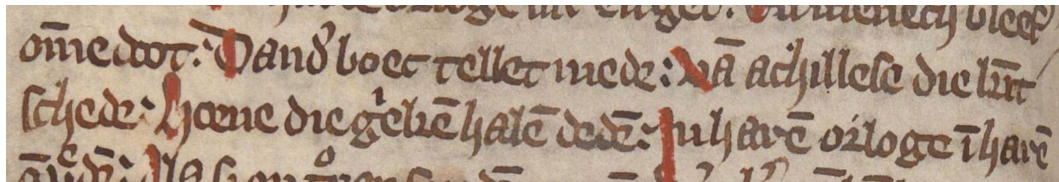
A major challenge of the Herne corpus are the aforementioned abbreviations, which are a hallmark of medieval handwritten cultures. Abbreviations (or brevigraphs [20]) were used to increase writing speed and conserve space. The Herne scribes adopted abbreviations profusely (especially in comparison to other contemporary texts in the vernacular), which they borrowed from Latin scribal cultures. Older editorial scholarship (e.g. in the Lachmannian tradition [36]) often aimed to reconstruct the authorial texts, rather than acknowledge individual copies as a scholarly object of study in their own right; in this paradigm, abbreviations were often treated as meaningless 'accidentals', that presented an obstacle, rather than an opportunity for research. Recent work in computational philology has rehabilitated the value of abbreviations, for instance, for scribal attributions [22, 17], linguistic analysis [38], stemmatology [1] and other innovative forms of material philology [12, 19, 30].

In this project too, we have adopted a (hyper-)diplomatic transcription standards that attempts to mimic the glyphs on the page, including the non-expanded abbreviations, as closely as possible. We report on an automated approach to expanding these abbreviations below. Throughout the project, we adhere to the diplomatic transcription MVN guidelines developed by Boot and Brinkman [3], which have been published online with a digital edition framework for the (popular, but proprietary) XML editing software *Oxygen* [url]. An illustration of this fairly deep and complex encoding in practice can be found in Fig. 3a. These digital guidelines are derived from those for the long-standing print edition series *Middeleeuwse Verzamelhandschriften uit de Nederlanden*, originally developed by Mertens [29]. Table 3 offers a selection of the most common abbreviation practices in the corpus, together with the character reference used in the MVN-guidelines to identify glyphs, as well as a number of illustrative examples. This table excludes multiple more specific, yet not uncommon brevigraphs, such as 'm^e' and 'm^n' for 'm*inne*' and ("love") 'm*innen*' ("to love") respectively or the dense logogram '$\bar{x}\bar{p}\bar{c}$' for '*christus*'.

Table 3

Common abbreviation practices in the corpus (selection) and their XML encoding. '<vow.>' = vowel.

MVN char. ref. (XML)	Expansion(s)	Examples (*abbrev.*)	Examples (*expand.*)
#bar	n – m – de	hē, gheuē, eñ	he*m*, gheue*n*, en*de*
#apomod	<vow.>+r	h'te, p'tie	he*r*te, pa*r*tie
superscript vowel	r+<vow.> – u+<vow.>	t^{e}ct, d^{a}get, qu^{a}de	t*re*ct, q*ua*de, d*ra*get
superscript t	eit, iet	n^t, domh^t	n*iet*, domh*eit*
#pbardes	per	dorꝑ, onghetempꝑ	dor*per*, onghetem*per*
#pcondes	con	ꝯde, onꝯnt	*con*de, on*cont*
#etfin	at – et – it	hꝫ, dꝫ, raꝫ	h*et*, d*at*, ra*et*
#pflour	pro	ꝓeue, ꝓper	*pro*eue, *pro*per
#usmod	us	Tulli^9, d^9	Tulli*us*, d*us*
#rrot	ur	nat^{2}e, auont^{2}e,	nat*ure*, auont*ure*

(a) Cutout from fol. 39v. (see Fig. 1) from the Heber-Serrure codex: lines 9-10 (Ghent, UL, 1374).

(b) XML snippet of the manual transcription of the lines above, following the MVN guidelines. Note how abbreviated tokens are represented using the `choice` element, which has an `abbr` and `expan` subnode. Under the `abbr` node, abbreviation glyphs are e.g. encoded using a g-node with a `ref` attribute; the expanded counterparts of such glyphs under the `expan` node are enclosed by `ex` tags.

Figure 3: Illustration of the XML encoding in a manually transcription, following the MVN guidelines.

4. Handwritten Text Recognition

For the transcription of our corpus, we leveraged the technology of Handwritten Text Recognition (HTR), specifically, the Transkribus platform [21, 31]. Initially, the digital images of all manuscripts in the corpus (see Table 1) were uploaded to this platform. An essential subsequent step in any HTR application involves layout analysis. This process involves identifying the specific region on each folio, demarcated by pixel coordinates, that contains the text. Additionally, the recognition of baselines within these designated regions is required, which determine the location of the text.

In order to prevent the misinterpretation of elements on the parchment surface as textual regions, these regions were manually outlined. Unlike the uniformity of blank paper, parchment often contains organic elements like hair follicles, gaps, and various other irregularities, which an automated layout model might mistakenly interpret as text regions. By manually outlining these regions, we managed to bypass this issue. Subsequent to the manual outlining of text regions, baselines within these regions were identified. For this procedure, we used an automated approach. Transkribus offers the possibility of training dedicated models specifically for recognizing baselines. For the task of baseline detection, we used two different models. One large baseline detection model was trained on a diverse collection of folia, for which the baselines were manually added. This model obtained an error rate of 7.00% on the training set consisting of 1,007 folia, and 6.35% on the validation set with 111 folia. In certain cases, it was beneficial to train dedicated baseline models, especially for manuscripts with unique layouts, such as VIENNA, ÖNB, COD. 13.708 and SAINT PETERSBURG, BAN, O 256. The former manuscript comprises closely packed dual-column text, while the latter contains numerous text lines blurred by water damage. Regardless of whether the general or dedicated model was employed, all baselines were manually reviewed. This process involved eliminating incorrectly recognized baselines and adjusting the length of overly short or long baselines. While this phase was the most labor-intensive in building the dataset, we consider it vital to assure the quality of the layout analysis, as it forms the cornerstone for all subsequent steps in the process.

Next, we produced sizable sample transcriptions for almost every document in the corpus using Transkribus. The transcription standard we adhered to, following the granularity levels of transcription described by Robinson and Solopova, can be characterized as 'graphemic reproductions'. This means that the spelling in the manuscripts is retained as is, and all abbreviations are transcribed exactly as they appear on the page. However, we made no distinction between letter forms (e.g., a long 'f' and a round 's' are both transcribed as 's').[1] We followed the Medieval Unicode Font Initiative (MUFI) for encoding and displaying special characters in medieval texts [16]. Furthermore, we consistently opted for the precomposed Unicode entities, ensuring that combinations of a letter character and a diacritical mark are always treated as a single entity.

A significant advantage for our project was the ability to bootstrap our transcriptions from certain existing editions, such as the manuscripts GHENT, UL, 1374 [14] and BRUSSELS, RL, 3093-95 (see below). We also utilized diplomatic editions of substantial sections of the manuscripts BRUSSELS, RL, 1805-8, BRUSSELS, RL, KBR 2877-78, and BRUSSELS, RL, 2879-2880 [22, 17]. These previously existing transcriptions were incorporated into the recognized baselines in Transkribus, using the text-to-baseline method, which is unfortunately currently deprecated. Following this, large segments of each manuscript were manually transcribed, guaranteeing that a minimum of 10% of the folia of each manuscript were manually transcribed. This approach ensured that our sample was representative of all manuscripts in our corpus.

In total, our full corpus contains 6,159 folios (both recto and verso-sides) of Middle Dutch text. After the manual transcription phase, we have 1,331 folios of Ground Truth text. The next step

---

[1]This is in line with the tradition in Middle Dutch scholarship [29], where scholars do not normally distinguish between the various allographs of <r> and <s>. We do distinguish, on the other hand, between <u> and <v> and do not normalize such instances, even if the medieval use of a glyph doesn't match the modern use.

Table 4
Training and validation results for the Middle Dutch HTR-Model.

Training of Middle Dutch HTR-model			
Train	1,197 folia	≈ 285,000 words	2.00 CER
Validation	133 folia	≈ 33,000 words	2.70 CER
Application of Middle Dutch HTR-model			
Remaining	4,828 folia	≈ 1,150,000 words	

was to use these materials to train a "Grand" Middle Dutch model to automatically transcribe the remaining 4,828 folios. A 90%-10% train-test split resulted in 1,197 folios for training and 133 folios for validation. By pooling together all this data and training this comprehensive model, we achieved a Character Error Rate (CER) of 2.7% on the validation set (see Table 4). This means that, on average, 2.7 out of every 100 characters are still incorrectly recognized when comparing the HTR output to the ground truth transcriptions. This is an encouraging result, bolstering our confidence in the model's capability to automatically transcribe various scribal hands, layouts, thematic contents, and spelling profiles.

5. Abbreviation resolution

To analyze the Herne manuscripts using conventional text technologies, we need an automated way to expand the abbreviations in the output of the HTR. Many brevigraphs are ambiguous, however, and can be expanded in multiple ways, depending on the context in which they occur (see Table 3). We therefore approach abbreviation resolution as a sequence modelling task in string transduction, in which an abbreviated line (or 'sentence') from a manuscript acquired through HTR is converted into a line of normalized tokens with their correctly expanded spelling. We operationalize this task at the token level, but need a model that can benefit from contextual information from surrounding tokens to resolve ambiguous instances. In essence, this task bears similarity to the traditional problem of lemmatization in the field of natural language processing. Following similar work by Camps et al. [5, 6], we resort to the PIE lemmatizer, which is an efficient lemmatizer for historic languages that abound in orthographic variation [28]. Crucially, this lemmatizer doesn't adopt a classification approach but uses recurrent neural architectures to generate a lemma, on a character-by-character basis. This means that it can also transduce tokens into normalized forms which it hasn't encountered during training.

5.1. Abbreviation materials

Our resources include (partial) transcriptions of four codices where the abbreviations have been manually resolved: (1) GHENT, UL, 1374 (Heber-Serrure MS by α; rhyming texts, but written down in continuous prose with inter-verse punctuation) [14]; (2) BRUSSELS, RL, KBR 2877-78 (*Letters* in Hadewijch MS B, scribe 10; prose formatted in two-column format) [22]; (3) BRUSSELS, RL, 2879-2880 (*Letters* in Hadewijch MS A; prose penned by multiple non-local scribes,

presented in a two-column format) [22]; and (4) Brussels, RL, 3093-95 (miscellany; complete transcription; multiple scribes; single-column prose).[2] In addition to the Herne material, we possess of transcriptions of multiple parallel versions of the thirteenth-century *Martijn*-trilogy by Jacob van Maerlant, which have been produced by dra. Sofie Moors (University of Antwerp) adhering to the same MVN guidelines and markup [url].[3] To establish how well the presented abbreviation solver scales to other material, with a diversity of scribal hands and dialects, we will also evaluate our procedure on this out-of-sample material. It should be noted, however, that the *Martijn* tradition, like most Middle Dutch texts, generally comprises far fewer abbreviations.

Below we report a series of experiments using a conventional setup, dividing the accessible material into training, validation, and test splits. Additionally, we report on a series of cross-domain and cross-scribal setups to stress-test the generalizability of our approach. We train PIE only for the lemmatization task, using the default configuration setups, but a relatively smaller layer size (to avoid overfitting, in what is clearly a less complicated task than full lemmatization). In Table 5, a random sample of pairs of abbreviated and expanded verse lines is provided. Given our focus on the line level, certain lines may begin or end with an incomplete token, a common occurrence in prose texts when words are split over two lines. Some lines need no changes at all, especially in the *Martijn* subset. Importantly, the lemmatizer shouldn't be greedy and must learn when to simply reproduce the input as output without alterations.

Table 6 displays descriptive summary statistics about the data per witness type: the number of available line pairs, the number of identical pairs, the length of the abbreviated/expanded lines (in characters), as well as the length difference between the two. A number of observations are important. The number of lines per Herne witness is fairly well balanced (25,017 lines in total), but the cumulative number of *Martijn* lines is much larger (note, however, that these are all parallel copies of the same text). The abbreviation rate is also the lowest in this witness category and more than half of the lines in fact need no abbreviation resolution at all, with very low mean difference scores overall. The manuscripts which presumably functioned locally in Herne (Ghent, UL, MSS 1374 and Brussels, RL, 3093-3095) clearly have the highest abbreviation density overall. The line length can also vary significantly, depending on the layout of a manuscript page. MS 1374, for instance, with its extravagantly dense prose layout, combines a very long line length with a high rate of brevigraphs: with some exaggeration, it would appear nearly *all* words in this codex are abbreviated.

The evaluation is reported in Table 6. The results are high for the in-domain test set, which was to be expected since there were no unknown tokens or targets. The few errors that persist are confusions between ambiguous resolutions; e.g. *-er* vs *-aer* for #apomod) or *-m* vs *-n* for #bar. In many cases, the limited context of a single line likely prevented a correct disambiguation. Only in a handful of cases, the system makes less forgivable errors, where it incorrectly copies a substring of the input, for instance. Some of the errors in the *Martijn* texts point to abbreviatory practices that are not encountered in Herne, such as #usmod for *-as* (instead of *-us*).

---

[2]This last manuscript has been transcribed by Dr. Ine Kiekens in 2018.

[3]This corpus in fact includes two copies that originated from Herne – both by α – which will be excluded from the *Martijn* corpus below.

Table 5

Random sample of 10 pairs of abbreviated and expanded lines in the manually transcribed corpus.

Abbreviated	Expanded	Witness
ests boetscap . Dan gheuȝ Mᵉ	ests boetscap . Dan gheuet Minne	Brussels, RL, 2879-80
Ghemaect eñ .i. ād' makē : Al ne warē nᵗ die sakē : Alsoe	Ghemaect ende .i. ander maken : Al ne waren niet die saken : Alsoe	Ghent, UL, 1374
dȝ hi heme : V'swo' . dād' alsict v'neme : Was ī ou'spele gᵒt :	dat hi heme : Verswoer . dander alsict verneme : Was in ouerspele grot :	Ghent, UL, 1374
Wes warachtich ende stout	Wes warachtich ende stout	*Martijn* (D)
eñ ōd' dē dorē : Dȝ seit die ewāgeliste wel : Om	ende onder den doren : Dat seit die ewangeliste wel : Om	Ghent, UL, 1374
eñ wōd'leke liedekene w'-	ende wonderleke liedekene wer-	Brussels, RL, 3093-3095
Ende welcker du liets sneuen/	Ende welcker du liets sneuen/	*Martijn* (D)
Dus ontbant my eñ beschiet.	Dus ontbant my ende beschiet.	*Martijn* (D)
begh'te altoes eñ hare	begherte altoes ende hare	Brussels, RL, 2877-78
Die alre meeste claerhᵗ die	Die alre meeste claerheit die	Brussels, RL, 2879-80

Table 6

Descriptive statistics on the abbreviation dataset per witness type. Columns represent: number of parallel lines, the proportion of identical pairs, the length of the abbreviated and expanded lines, and the difference between the length of the expanded and abbreviated lines. (Line length measurements are expressed in UTF-8 character counts, so that these counts are in fact underestimations, because glyphs like the tilde don't contribute to the length of the running texts.)

Witness	Pairs	Identical	Abbreviated	Expanded	Difference
Brussels, RL, 2879-80	5703	0.26	26.02	27.91	1.89
Brussels, RL, 3093-3095	6017	0.04	30.46	34.46	4.00
Brussels, RL, MS 2877-78	6428	0.20	22.36	24.27	1.92
Ghent, UL, 1374	6869	0.02	46.55	51.96	5.41
Martijn	13924	0.55	25.71	26.60	0.89

6. Linguistic annotation

Following the resolution of abbreviations in the original transcriptions, standard language technology can be applied to the corpus. (Middle) Dutch studies have been blessed with rich lexicographic resources developed by the Dutch Language Institute. We have applied a recently developed lemmatizer and part-of-speech tagger on the basis of the PIE software [28], which was trained on all of the available resources for Middle Dutch, including a subcorpus of religious and mystical writings that should suit the Herne material well [23]. The used model (DUM) is publicly available from a framework developed by Clérice [9] that makes it easy to apply PIE models to new data. This linguistic enrichment layer allows scholars to abstract over surface variation, which can be beneficial for studies that wish to abstract over scribal accidentals (for instance, in the context of search), but also for applications in stylometry or

Table 7

Results for three abbreviation resolution experiments, in terms of accuracy, F1 score and number of test items (support): *Herne*: random test split (12.5%) on all Herne MSS; *Martijn*: performance on out-of-sample texts; *Out-house*: training on MSS that were meant to function inside the charterhouse vs testing on MSS that weren't meant as local products.

	Herne			*Martijn*			*Out-house*		
	acc	**F1**	**sup**	**acc**	**F1**	**sup**	**acc**	**F1**	**sup**
all tokens	.996	.970	132,518	.994	.970	21,966	.991	.961	21,966
ambiguous tokens	.967	.485	8,211	.974	.747	1,346	.947	.574	1,378
known tokens	.996	.970	132,518	.997	.984	20,062	.994	.976	20,179
unknown targets	–	–	–	.964	.931	1,526	.952	.912	1,364
unknown tokens	–	–	–	.970	.942	1,904	.954	.919	1,787

semantic analysis.

In conjunction with this paper, all related materials are released on the Zenodo data repository platform, accessible via the following link. These materials include the transcriptions in an XML scheme, metadata, and linguistic annotation layers. All these materials are available under a CC-BY 4.0 licence (Attribution 4.0 International). Users are allowed to share and adapt these resources for any purpose, with the condition of providing adequate attribution to this paper. However, due to unresolved intellectual rights (refer to the links in Table 1), the original images will be stored in a separate "restricted access" Zenodo repository. They can only be shared within the context of a scientific collaboration under Belgian copyright law. Finally, the corpus-wide *Transkribus* HTR model is set to be made publicly available on the software platform.

7. Discussion: an un-edition?

In this dataset paper, we reported on the construction of a near-exhaustive collection of digital transcriptions of the Herne manuscripts that is readily and openly available for computational text analysis. Because the project's focus lies with studying Herne as a scribal community, we have chosen to produce a hyper-diplomatic rendering of the manuscripts. Apart from theoretical considerations, this choice was also meant to help the HTR engine, because the target transcriptions stay relatively closer to the glyphs on the page and, thus, require less contextual interpretation and disambiguation. These choices set us apart from most of the prior Herne editions that have been heavily text critical in orientation [4, 2, 7]. Older editions abstracted over various manuscript versions of a text, in an attempt to reconstruct the original authorial version. Critical editions typically normalize a text's surface variation [33]: editors adapt spellings (e.g. allographs) to modern orthographic conventions and they resolve abbreviations in a manner that doesn't allow to reconstruct the original glyphs on the page. For all these reasons, critical editions do not optimally suit the computational analysis of scribal profiles.

In a recent online seminar,[4] the traditional endeavour of the critical, closed, printed edition

---

[4] *The Birth of the UnEdition* led by Ben Albritton and Laura Morreale under the umbrella of the 'DarkArchives 20/21'

Table 8
Comparing editions to "un-editions"

	Edition	Un-Edition
	Printed (in paper)	Digital (in the cloud)
	Printed (in paper)	Collaborative (asynchronous)
	Static (for eternity)	Dynamic (until next bugfix)
	Critical (multi-witness)	Hyper-diplomatic (single-witness)
	Hide sources	Foreground sources
	Old philology	New philology
	Accuracy	Accessibility
	Paywalled, copyrighted	Open science
	Focus on text	Focus on document

was put in a stimulating, yet perhaps somewhat undue opposition with the novel forms of collaborative, open, digital edition – tentatively termed an "un-edition". In Table 8 we survey some of the oppositions that were discussed at this event. Exaggerated by the global lockdown conditions during the COVID pandemic, the digital medium has increased the need among medievalists for digital editions that are widely and openly accessible, beyond the conventional consultation in physical libraries. Increasingly, editions are a collaborative, asynchronous endeavour. Static, printed editions, published for eternity, lose their relevance in the digital sphere: apart from the paywalls that limit the accessibility, un-editions have the advantage of being dynamic – they are open to constant improvement. Because of the ubiquity of digital facsimiles (that are increasingly served over open protocols like IIIF) as well as the affordances of handwritten text recognition, the diplomatic edition is going through a clear renaissance. In the words of Fisher in a thought-provoking article:

> 'The provision of digital facsimiles has put the manuscripts, the very material base of any editorial endeavor, into focus again. Several editions have been created that engage primarily with individual manuscripts; others have posited a wide range of variance as a central characteristic of medieval literature instead of relegating variants to the footnotes of ahistorically normalized and regularized texts or speculative reconstructions of archetypes and authorities' [13, p. S265].

Fisher goes on to stress that the idea of a critical edition is not obsolete – on the contrary, in his view the need for critical editions has only increased with the availability of document-oriented transcriptions: 'without a critical text the modern reader is at a loss to adjudicate on the quality of the textual version picked up randomly on the internet' (*ibid.*). He nevertheless notes at the same time that only a relatively small number of born-digital critical editions of texts exists, although the necessary technology is in fact available. This observation touches upon a more fundamental development in medieval studies, where scholars have come to prefer availability over accuracy in recent years. This aspect is also crucial to the concept of an un-edition, which favors an openly available text over a more accurate, yet closed text. This might

---

inter-conference, on Thursday 18th March 2021, via Zoom. See [url]

be a controversial point but colleagues seem, overall, quite happy to strike a scholarly trade-off between the poor textual quality of mega-collections like the Internet Archive or Google Books and the vast serendipity which these textual reservoirs enable. The COVID pandemic has, in this sense, fast-tracked scholarly developments that had been emerging over a much longer period, in the wake of the material or "artefactual" philology [37]. Currently, many terms have surfaced to describe the new forms of digital editorial scholarship: un-edition, proto-edition, artificial edition [34].

In conclusion, our project offers a fresh approach to the study of the Herne manuscripts, emphasizing the usefulness of hyper-diplomatic transcriptions and leveraging digital technology to render them accessible to a broader scholarly community. The concept of the "un-edition", although still emerging and somewhat controversial, holds promise as it aligns with the ongoing shift in medieval studies from accuracy to availability. By making the original, abbreviated transcriptions of our corpus available online, along with the fully expanded and lemmatized texts, part-of-speech tags, and extensive metadata, we embrace this transformation. In this sense, we hope our work underscores the dynamic nature of digital humanities, which, much like the manuscripts we study, are in constant flux.

Acknowledgments

References

[1] T. Andrews and C. Macé. "Beyond the Tree of Texts: Building an Empirical Model of Scribal Variation through Graph Analysis of Texts and Stemmata". In: *Literary and Linguistic Computing* 28.4 (2013), pp. 504–521. DOI: 10.1093/llc/fqt032.

[2] A. Berteloot, G. Claassens, and W. Kuiper, eds. *Gulden Legende: De Middelnederlandse vertaling van de 'Legenda aurea' door Petrus Naghel, uitgegeven naar handschrift Brussel, Koninklijke Bibliotheek, 15140. Deel II.* Brepols, 2012.

[3] P. Boot and H. Brinkman. *Richtlijnen voor digitale diplomatische edities in de reeks* Middeleeuwse Verzamelhandschriften uit de Nederlanden. 2020. URL: https://github.com/HuygensING/mvn-xml/blob/main/docu/Richtlijnen%5C%20MVN%5C%20digitaal.pdf.

[4] K. D. Bundel. "*Van woerde tot woerde oft van synne te sinne.* Petrus Naghel en het translatorium van de kartuis te Herne (ca. 1350-1400)". PhD thesis. KU Leuven, 2009.

[5] J.-B. Camps, T. Clérice, and A. Pinche. "Noisy medieval data, from digitized manuscript to stylometric analysis: Evaluating Paul Meyer's hagiographic hypothesis". In: *Digital Scholarship in the Humanities* 36.Supplement_2 (2021), pp. ii49–ii71. DOI: 10.1093/llc/fqab033.

[6] J.-B. Camps, C. Vidal-Gorène, and M. Vernet. "Handling Heavily Abbreviated Manuscripts: HTR Engines vs Text Normalisation Approaches". In: *Document Analysis and Recognition – ICDAR 2021 Workshops*. Ed. by E. H. Barney Smith and U. Pal. Cham: Springer International Publishing, 2021, pp. 306–316.

[7] G. Claassens, W. Kuiper, and A. Berteloot, eds. *Gulden legende. De Middelnederlandse vertaling van de Legenda aurea door Petrus Naghel*. Brepols, 2017.

[8] G. H. M. Claassens and K. de Bundel. "Petrus Naghel, Übersetzer in Herne". In: *Zeitschrift Für Deutsche Philologie* 130 (2011), pp. 267–281.

[9] T. Clérice. *Pie Extended, an extension for Pie with pre-processing and post-processing*. 2020. DOI: 10.5281/zenodo.3883589. URL: https://doi.org/10.5281/zenodo.3883589.

[10] T. Coun. "The Bible Translator of 1360". In: *Ons Geestelijk Erf* 85.1-2 (2014), pp. 3–38.

[11] J. Deschamps, ed. *The Vienna manuscript of the "Second part" of the "Spiegel historiael": Ms. Vienna, Österreichische Nationalbibliothek, 13.708*. Medieval manuscripts from the Low Countries in facsimile. Rosenkilke and Bagger, 1971.

[12] M. J. Driscoll. "The Words on the Page: Thoughts on Philology, Old and New". In: *Creating the Medieval Saga: Versions, Variability, and Editorial Interpretations of Old Norse Saga Literature*. Ed. by J. Quinn and E. Lethbridge. Odense: University of South Denmark, 2010, pp. 87–104.

[13] F. Fischer. "Digital Corpora and Scholarly Editions of Latin Texts: Features and Requirements of Textual Criticism". In: *Speculum* 92.S1 (2017), S265–s287.

[14] R. Gabriël and M. Kestemont, eds. *De Heber-Serrurecodex: Gent, Universiteitsbibliotheek, Hs. 1374. Diplomatische editie bezorgd door Renée Gabriël en Mike Kestemont, met een dialectologische analyse door Amand Berteloot*. Middeleeuwse Verzamelhandschriften uit de Nederlanden XVII. Amsterdam: Huygens Instituut voor Nederlandse Geschiedenis en Cultuur van de Koninklijke Nederlandse Akademie van Wetenschappen, 2023. DOI: 10.5281/zenodo.8385501. URL: https://hbsr.mvn.huygens.knaw.nl.

[15] T. Gaens and J. de Grauwe. *De kracht van de stilte. Geest & geschiedenis van de kartuizerorde*. Peeters Publishers, 2006.

[16] O. E. Haugen. *MUFI character recommendation v. 3.0*. 2009.

[17] W. Haverals and M. Kestemont. "From exemplar to copy: the scribal appropriation of a Hadewijch manuscript computationally explored". In: *Journal of Data Mining & Digital Humanities* On the Way to the Future of Digital Manuscript Studies (2023). DOI: 10.46298/jdmdh.10206.

[18] W. Haverals and M. Kestemont. "Silent voices. A Digital Study of the Herne Charterhouse Scribal Community (ca. 1350-1400)". In: *Queeste* 27.2 (2020), pp. 186–195. DOI: 10.5117/que2020.2.006.have.

[19] A. Honkapohja. "Digital Approaches to Manuscript Abbreviations: Where Are We at the Beginning of the 2020s?" In: *Digital Medievalist* 14 (2021). DOI: 10.16995/dm.88.

[20] K. Honkapohja. "Digital Approaches to Manuscript Abbreviations: Where Are We Now?" In: *Digital Medievalist* 5.1 (2009). DOI: 10.16995/dm.88.

[21] P. Kahle, S. Colutto, G. Hackl, and G. Mühlberger. "Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents". In: *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 4. 2017, pp. 19–24.

[22] M. Kestemont. "A computational analysis of the scribal profiles in two of the oldest manuscripts of Hadewijch's letters". In: *Scriptorium* 69.1 (2015), pp. 159–177. DOI: 10.1484/j.scri.5.103464.

[23] M. Kestemont, G. de Pauw, R. van Nie, and W. Daelemans. "Lemmatization for variation-rich languages using deep learning". In: *Digital Scholarship in the Humanities* 32.4 (2016), pp. 797–815. DOI: 10.1093/llc/fqw034.

[24] M. Kors. *De bijbel voor leken. Studies over Petrus Naghel en de historiebijbel van 1361*. Brepols, 2007.

[25] E. Kwakkel. "A meadow without flowers. What happened to the Middle Dutch manuscripts from the Charterhouse Herne?" In: *Quaerendo* 33.1-2 (2003), pp. 191–211.

[26] E. Kwakkel. *Die Dietsche boeke die ons toebehoeren. De kartuizers van Herne en de productie van Middelnederlandse handschriften in de regio Brussel (1350-1400)*. Peeters, 2002.

[27] E. Kwakkel. "Towards a terminology for the analysis of composite manuscripts". In: *Gazette du livre médiéval* 41 (2002), pp. 12–19.

[28] E. Manjavacas, Á. Kádár, and M. Kestemont. "Improving Lemmatization of Non-Standard Languages with Joint Learning". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 1493–1503. DOI: 10.18653/v1/N19-1153.

[29] T. Mertens. *Richtlijnen voor de uitgave van Middeleeuwse Verzamelhandschriften uit de Nederlanden*. Hilversum: Verloren, 1994.

[30] S. Moors. "Een bijzondere Brabantse tekstgetuige van de Martijntrilogie van Jacob van Maerlant. De fragmenten Lyon, Bibliothèque municipale, ms 6848". In: *Queeste* 29.1 (2022), pp. 36–65. DOI: 10.5117/que2022.1.002.moor.

[31] G. Muehlberger, L. Seaward, M. Terras, S. Ares Oliveira, V. Bosch, M. Bryan, S. Colutto, H. Déjean, M. Diem, S. Fiel, B. Gatos, A. Greinoecker, T. Grüning, G. Hackl, V. Haukkovaara, G. Heyer, L. Hirvonen, T. Hodel, M. Jokinen, P. Kahle, M. Kallio, F. Kaplan, F. Kleber, R. Labahn, E. M. Lang, S. Laube, G. Leifert, G. Louloudis, R. McNicholl, J.-L. Meunier, J. Michael, E. Mühlbauer, N. Philipp, I. Pratikakis, J. Puigcerver Pérez, H. Putz, G. Retsinas, V. Romero, R. Sablatnig, J. A. Sánchez, P. Schofield, G. Sfikas, C. Sieber, N. Stamatopoulos, T. Strauß, T. Terbul, A. H. Toselli, B. Ulreich, M. Villegas, E. Vidal, J. Walcher, M. Weidemann, H. Wurster, and K. Zagoris. "Transforming scholarship in the archives through

handwritten text recognition: Transkribus as a case study". In: *Journal of Documentation* 75.5 (2019), pp. 954–976.

[32] F. van Oostrom. *Wereld in woorden: Geschiedenis van de Nederlandse literatuur 1300-1400.* Amsterdam: Prometheus, 2013.

[33] E. Pierazzo. *Digital Scholarly Editing: Theories, Models and Methods.* Routledge, 2015.

[34] *Proto-editions: Historians and the "Something between digital image and digital scholarly edition".* 2023. DOI: 10.5281/zenodo.8107922. URL: https://doi.org/10.5281/zenodo.8107922.

[35] P. Robinson and E. Solopova. "Guidelines for transcription of the manuscripts of the Wife of Bath's Prologue". In: *The Canterbury Tales Project. Occasional papers* I (1993).

[36] P. Roelli, ed. *Handbook of Stemmatology. History, Methodology, Digital Approaches.* Berlin, Boston: De Gruyter, 2020. DOI: 10.1515/9783110684384.

[37] L. Rohrbach. "Material Philology". In: *Handbook of Pre-Modern Nordic Memory Studies. Interdisciplinary Approaches.* Ed. by J. Glauser, P. Hermann, and S. A. Mitchell. Berlin, Boston: De Gruyter, 2019, pp. 210–216. DOI: 10.1515/9783110431360-020.

[38] J. Thaisen. "Standardisation, exemplars, and the Auchinleck manuscript". In: *The Multilingual Origins of Standard English.* Ed. by L. Wright. Berlin, Boston: De Gruyter Mouton, 2020, pp. 165–190. DOI: 10.1515/9783110687545-006.

[39] G. Warnar. *Geert Ruusbroec: Literatuur En Mystiek In De Veertiende Eeuw.* Amsterdam: Athenaeum-Polak & Van Gennep, 2003.