

Registered Report Protocol

Cite this article: J. Maliszewski 2025. "It takes a village to write a book: Mapping anonymous contributions in Stephen Langton's *Quaestiones Theologiae*. *Computational Humanities Research*, 1:e2, <https://doi.org/10.1017/chr.2025.10004>

Received: 4 March 2025

Revised: 15 April 2025

Accepted: 26 May 2025

Keywords:

handwritten text recognition; reportationes; scholasticism; Stephen Langton; stylometry

It takes a village to write a book: Mapping anonymous contributions in Stephen Langton's *Quaestiones Theologiae*

Jan Maliszewski 

Faculty of Philosophy, University of Warsaw, Warszawa, Mazowieckie Voivodeship, Poland

Email: j.maliszewski@uw.edu.pl

Abstract

While the indirect evidence suggests that already in the early scholastic period the literary production based on records of oral teaching (so-called *reportationes*) was not uncommon, there are very few sources commenting on the practice. This article details the design of a study applying stylometric techniques of authorship attribution to a collection developed from *reportationes* – Stephen Langton's *Quaestiones Theologiae* – aiming to uncover layers of editorial work and thus validate some hypotheses regarding the collection's formation. Following Camps, Clérice, and Pinche (2021), I discuss the implementation of an HTR pipeline and stylometric analysis based on the most frequent words, POS tags, and pseudo-affixes. The proposed study will offer two methodological gains relevant to computational research on the scholastic tradition: it will directly compare performance on manually composed and automatically extracted data, and it will test the validity of transformer-based OCR and automated transcription alignment for workflows applied to scholastic Latin corpora. If successful, this study will provide an easily reusable template for the exploratory analysis of collaborative literary production stemming from medieval universities.

Plain language summary

Many texts produced at the medieval universities did not originate as literary works but were instead gradually and collaboratively developed from records of oral teaching, known as *reportationes*. While this practice was likely widespread, there are very few sources detailing its daily operation, forcing scholars to rely on indirect evidence deducible from preserved works. In this context, this article proposes a study exploring computational analysis of style as a way to track layers of editorial work in scholastic collections, potentially revealing the actual scope of authors' control over these works. This approach draws from earlier studies which successfully employed computational techniques in the context of medieval Latin letter collections and Old French hagiographies. I discuss applying similar methods to the collection of Stephen Langton's (d. 1228) theological *quaestiones*. Langton's collection is particularly interesting for it is known to depend on *reportationes*, and it transmits most of its material in more than one version, in some cases allowing us to track the development from raw records of oral teaching to fully developed literary forms. Initial analysis of Langton's corpus shows that by measuring the frequencies of the most common words – a common stylometric method – it is possible to differentiate its stylistic signal from other contemporary scholastic collections, as well as to observe some stylistic diversity within Langton's corpus. However, the key limitation in the context of Langton's *quaestiones* stems from their length, as most of *quaestiones* are too short to provide representative samples. This issue can be addressed by including additional stylistic features: sequences of Part of Speech tags, which capture syntactic structures, and pseudo-affixes (the few opening and closing characters of each word), which represent morphological information. These features have been shown to provide good results with automatically generated transcriptions; consequently, I plan to compare tests performed on manually composed editions and automatically extracted data. The key gain offered by automated transcription lies in providing a feasible way of extending analysed corpora by including unedited material.

Introduction

This article proposes a study employing stylometric techniques of authorship attribution to assess the scope of anonymous contributions to the collection of Stephen Langton's *Quaestiones Theologiae*. In this, it follows studies which demonstrated the robustness of stylometric methods applied to the analysis of collaborative authorship in comparable medieval Latin literary traditions (Kestemont, Moens, and Deploige 2013; De Gussem 2017). In particular, I draw heavily on the methods of unsupervised cluster analysis offered in Camps and Cafiero (2013), Cafiero and Camps (2019), Camps, Clérice, and Pinche (2021). The central goal of the proposed

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial licence (<http://creativecommons.org/licenses/by-nc/4.0>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original article is properly cited. The written permission of Cambridge University Press must be obtained prior to any commercial use.



study is to analyse stylistic signals observable within a collection known to originate from anonymous *reportationes* – the collection of Stephen Langton's *Quaestiones Theologiae* – aiming to locate any internal stylistic clusters. The hypothesis is that, if discernible, such clusters may be representative of the activity of non-authorial contributors. While the proposed study's design is informed by recent editorial work on Langton's collection (Langton, ed. Bieniak et al. 2014–2024), these methods can be expected to apply to other scholastic corpora displaying similar traces of collaborative work. To further explore this potential transfer of methods, the proposed study will involve a direct comparison of the performance of the stylometric tests on both manually edited and HTR-extracted data, adapting the pipeline constructed in Camps, Clérice, and Pinche (2021). Below, I discuss the philological motivation of the problem, followed by a discussion of the selected methods and potential results.

State of research on early scholastic reportationes

Dating back at least to the 1920s, the scholarly interest in the production of *reportationes* gradually led to their recognition as a salient feature of the scholastic intellectual practice.¹ Generally speaking, a *reportatio* is a note recording oral teaching, usually taken from a master's lecture by one of its participants. The proliferation of *reportationes* was closely associated with the growth of universities, and many attempts were made to analyse *reportationes* in the context of specifically medieval didactic forms. Thus, for example, *reportationes* prove uniquely valuable as testimonies of the practice of formal public debate, *disputatio*, in the 13th and 14th centuries.² Still, *reportatio* as such was neither a genre nor a transmission method but a technique applied in many different contexts and with varying aims.³ In many cases, the primary goal of such note-taking may have been private, intended to aid the student's memory. However, there are also documented cases in which the teaching collected through *reportationes* formed the foundation of a master's regular literary works. It is not always easy to establish whether a particular text originated from *reportationes*, and thus the scope of such oral-to-literary transfer is not fully understood. While the literary production based on *reportationes* dates back at least to the 1120s, for the entire 12th century scholars have identified only two testimonies describing the process of reporting and its later literary refinement.⁴ Consequently, the existing research on the earliest usage of *reportationes* for literary production – that is, the production stemming from the cathedral schools and universities before c. 1250 – largely extrapolates from these two testimonies and the more comprehensive information available for later scholastic tradition.

Two basic types of evidence provide insight into the actual scope of the early scholastic literary production based on *reportationes*. First, scholars identified marks of oral communication in some

otherwise inconspicuous literary works. These marks can be lexical or pragmatic. Examples include the prevalence of second-person verb forms, ellipses, or context-specific references to the audience – e.g., singling out lecture participants by name or recalling earlier exchanges of arguments, not preserved in the written testimony.⁵ Another type of indirect evidence is stemmatical. It is not uncommon for traditions dating back to 12th-century Paris to transmit multiple partially collatable versions, likely indicating independent strands of transmission in the text's early history. Transmission via *reportationes* is a likely cause behind at least some of this variance,⁶ especially when more than one record of a lecture was created and when the master did not supervise the process. Taken together, available evidence suggests that already in the early stages of the scholastic tradition, it was fairly common for a master to produce his works from *reportationes*.

Different general accounts of the practice of *reportatio* can be largely traced back to scholars' interest in corpora exhibiting different consequences of transmission via *reportationes*. Some collections, while demonstrably stemming from classroom reports, are stemmatically regular – that is, the stemmatical evidence suggests the existence of a single archetype at the origin of the tradition – leading their editors to assume a higher degree of reportatorial professionalisation and master's control over the process.⁷ On the other end of the spectrum, we find collections compiling and reworking scattered reportatorial material, possibly with little or no magisterial control, and at a considerable time distance from the initial lecture.⁸ Overall, the *reportatio* seems to be less of a formalised and unified phenomenon in the 12th century than in its later practice, and thus many basic questions relating to its operation remain open. In particular, in most cases we do not know how many actors – and with what exact roles – stand behind the preserved collections. A model transmission would involve the reportator reworking his record shortly after the class or debate, presumably mostly to supplement the details missing due to the hastiness of the initial record,⁹ and then the master authenticating the testimony, likely extensively interfering in the text – this final correction is known as an *ordinatio*.¹⁰ How closely the daily operation of textual production based on *reportationes* resembled this schema is not clear, but we can safely assume that the preserved records are skewed on the side of more regular instances of reporting, as these were more likely to enter into wider circulation requiring ample scribal work.

⁵For a comprehensive discussion of markers of orality preserved in 12th-century collections, see Siri (2013).

⁶Other likely factors shaping irregular transmission in this period include evolution of the text after its initial circulation – both authorized by the master and independent, e.g., by incorporation of external glosses – and transmission *per pecia*, i.e., the practice of copying long works from smaller booklets, which may have easily resulted in the circulation of incomplete witnesses. On *reportationes*, dictation, and the practice of transmission *per pecia* in medieval Paris, see Weijers (2015, 165–74).

⁷An example of such a regular 12th-century collection developed from *reportationes* can be found in Peter Comestor's Gospel glosses – see Peter Comestor, ed. Foley, 2024, especially the introductory discussion on pp. 17–20.

⁸This, as discussed below, is the case of Stephen Langton's *Quaestiones*.

⁹It should be noted that any preserved record is virtually never identical with the initial *reportatio* since, as far as we know, these were ordinarily produced on provisional writing support, e.g., wax tablets or loose offcuts of parchment. Moving such text to regular parchment folios likely involved at least a minimal degree of editorial normalisation.

¹⁰This is the process described by Laurentius, Hugh of St. Victor's pupil reporting *Sententiae de divinitate* (see the reference in n. 4 above).

¹For a historical summary of the literature on *reportationes*, see Saccenti (2016, 74–6).

²See Hamesse (1997, 420). For a comprehensive study of the practice of *disputatio*, see Weijers (2013).

³As commonly acknowledged after Hamesse (1997). A notable context outside of this study's scope is sermon *reportationes* – see Roberts (1968), d'Avray (1985), Bériou (2020).

⁴These testimonies come from Abelard's account of his exegetical lectures (Abelard, ed. Monfrin 1959, pp. 69–70), and from a letter of an otherwise unknown Laurentius, the *reportator* of Hugh of Saint-Victor's *Sententiae de divinitate* (Hugh of St. Victor, ed. Piazzoni, 1982, pp. 912–3). For discussion of these passages, see Siri (2013), Foley (2024, 16–29).

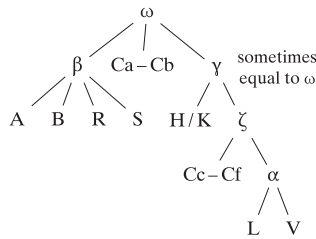


Figure 1. General transmission pattern of Langton's Quaestiones Theologiae.

Corpus: Stephen Langton's Quaestiones Theologiae

The collection of Stephen Langton's Quaestiones provides a particularly convenient vantage point for the study of the practice of *reportatio* in the early university setting. Stemming from Langton's Parisian teaching sometime during the last decades of the 12th century up to 1206, this collection was never given a final shape, despite some clear traces of attempted editorial work. Around 70% of the *quaestiones* listed in the contemporary index of the collection are transmitted in multiple substantially different versions, preserved at varying stages of production.¹¹ Some of these include exceptionally concise discussions – presumably unedited transcripts of *reportationes* – which correspond with some of the fully developed *quaestiones*, either preserving the structure of the argumentation or being partially collatable, suggesting that these versions represent different accounts of one oral *quaestio*.

The collection is transmitted by eight major manuscript witnesses (Figure 1). The discernible subcollections (mss. C,¹² H / K, and families α and β) likely represent parallel, partially overlapping compilations of Langton's material. They transmit vastly different sets of *quaestiones*, mostly in varying order. Part of the collection may have been reviewed by Langton – especially in ms. C – but most of the *quaestiones* were almost certainly edited by someone else, possibly by unknown students or secretaries from Langton's milieu after 1206. How many editors worked on this collection remains unclear. Similarly, we have no estimate of the number of *reportatores* involved in recording Langton's teaching.

Exploratory stylometric analysis

The basic premise of the proposed study stems from the results of Kestemont, Moens, and Deploige (2013) and De Gussem (2017). Both these studies applied techniques of stylometric authorship attribution in the context of 12th-century collaborative Latin writing, showing that it is possible to track with these tools stylistic variance which can be linked to the contributions of secretaries working with, respectively, Hildegard of Bingen and Bernard of Clairvaux. Our hypothesis – to some extent validated by the exploratory analysis – is that it is similarly possible to map the layers of reportatorial and editorial activity in scholastic corpora.

Both these studies employed to a good effect a widely accepted metric of style: the frequencies of function words, that is, the

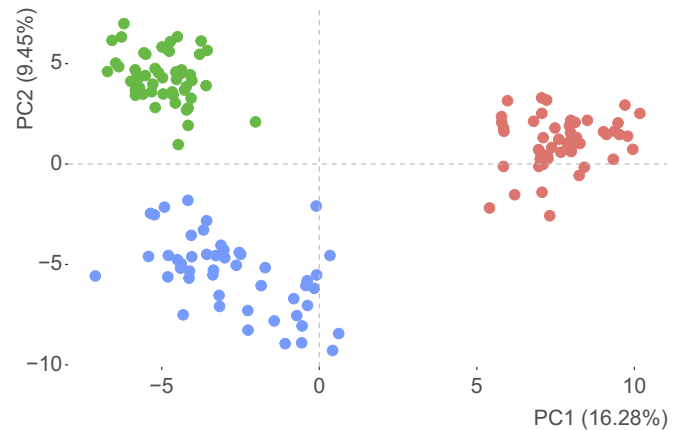


Figure 2. PCA of samples from Aquinas (red), Courson (green), and Langton (blue).

most common subject-independent lemmas observed in a given corpus.¹³ While, as discussed below, the specific stylometric tests applied in these studies do not transfer well into the problem at hand, it can certainly be confirmed that function words provide a reliable marker of style for scholastic corpora. For example, Figure 2 shows a comparison of 3,000-word samples from Langton's *quaestiones*, Robert of Courson's *Summa*,¹⁴ and Aquinas' *Summa Theologiae, prima pars*.¹⁵ From each text, we draw 50 continuous samples. All samples are represented by the relative frequencies of the 200 most frequent words (unlemmatized), which largely align with function words. The data was transformed by primary component analysis (PCA), with the two top components capturing a little over 25% of the total variance. As apparent in the plot, all samples cluster according to their text of origin, showing that these authorial signals can be identified based on the usage of the most frequent words. It is not surprising – function words prove to be effective across many languages and genres – but also not entirely trivial, since theological *quaestiones* of the period belong to a highly technical and formulaic genre, and thus can be expected to display overall fainter stylistic signals than the related epistolary or sermon corpora.

Two factors limit the usefulness of the above test for the analysis of stylistic clusters within Langton's collection. First, since we have no reliable estimate of the number of expected classes, PCA alone is not a suitable clustering mechanism, as it can conflate some clusters discernible in the initial data. The second limitation is related to the samples' length. For the distributions of the most frequent words to be representative of the authorial signal, the sample length needs to reach a threshold of 2,000–5,000 words, with the exact required length varying depending on genre and language (Eder 2013). Meanwhile, the average length of a single *quaestio* in Langton's collection is around 1,400 words, with the extreme values of 166 and 7,385 words.¹⁶ We can reach the reliable

¹³For example, the ten most frequent words (unlemmatized) in Langton's corpus are "est," "et," "non," "quod," "in," "ergo," "set," "ad," "quia," and "hoc."

¹⁴On Robert's *Summa*, see Kennedy (1947). I used a transcription of ms. Bruges 247, ff. 4^{va}–61^{va}, kindly shared by Gary Macy.

¹⁵*Summa Theologiae*, I^a, qq. 1–45, following the text of *Corpus Thomisticum*.

¹⁶The numbers reported here and in Table 1 are representative of all published or preliminarily edited *quaestiones*, which cover roughly 90% of the entire material. The ongoing critical edition of *quaestiones* (Langton, ed. Bieniak et al., 2014–2024) is planned for six volumes, four of which are already published, and the remaining two are at an advanced stage.

¹¹Of the 173 *quaestiones*, 119 are transmitted in two to five different versions. These numbers do not account for the so-called *quaestiones extra indicem*; including these texts and all the versions, the collection contains over 350 different texts. For the complex issues of cataloguing Langton's *quaestiones*, see Quinto (1994) and the introduction to the first volume of the critical edition of Langton's collection, ed. Quinto, Bieniak (2014).

¹²Ms. C consists of six distinct codicological units, Ca–Cf, which occupy different positions in the stemma.

Table 1. Grouping Langton's *Quaestiones* by shared codices.

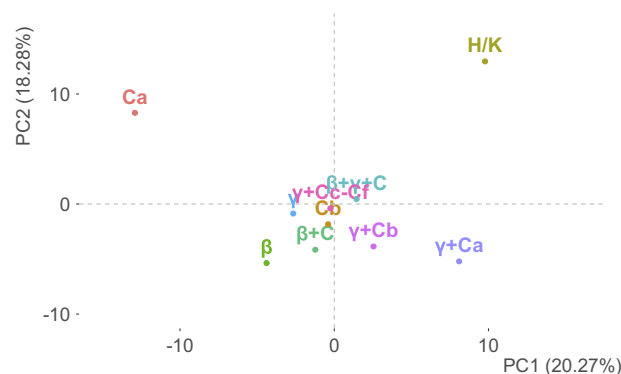
Class by transmitting mss.	No. of <i>quaestiones</i>	Total length (in words)
β	98	106,221
γ with Cc – Cf	65	86,543
$\beta + \gamma + C$ (any section)	32	76,215
Cb	54	70,282
$\gamma + Cb$	27	36,776
γ without C	23	27,113
$\beta + C$ (any section)	12	23,769
Ca	9	18,313
$\gamma + Ca$	9	17,251
H / K	11	11,125

Note: For more details on this data, consult the supplementary files – see the Data Availability Statement below.

sample's length by concatenating the *quaestiones* – as in the above test – but this effectively averages over the stylistic signal of all *quaestiones* included in a given sample, obscuring the signals of shorter texts and under-representing the actual stylistic variance of the collection.

While the most promising way to address this issue seems to be by extending the set of analysed features – see the discussion in the “Methods” section below – this problem can be to some extent mitigated by bundling the *quaestiones* according to the information obtained from stemmatical analysis. As already noted, the *quaestiones* are transmitted in four subcollections, which contain different, partially overlapping sets of *quaestiones*. Since these subcollections most likely originated as compilations of dispersed Langtonian material, it makes sense to analyse smaller classes of *quaestiones* organised by the set of manuscripts in which they are transmitted. In this way, we end up with 10 disjoint classes, as detailed in Table 1. This organisation of material accounts for major stemmatical relations, including the shifting relation between ms. C and family γ .¹⁷

Figure 3 shows the results of PCA conducted for these classes, based on the distribution of the 200 most frequent words. While most classes expectedly cluster around the average for the entire collection, there are two clear outliers: the material transmitted exclusively in section Ca of ms. C, as well as *quaestiones* proper to the Chartres collection H / K.¹⁸ In the case of Ca, this notably aligns with a long-standing palaeographic observation: the final folios of Ca – the ones transmitting material not found in any other codices – were copied by a different hand (Gregory 1930). Similarly, the bulk of *quaestiones* transmitted solely by H / K is positioned on its final folios (ms. K, f. 152ra–153va), possibly also

**Figure 3.** PCA for Langton's *Quaestiones*, grouped by transmitting codices.

copied *alia manu*.¹⁹ Thus, the exploratory analysis shows that even based on this admittedly unrefined set of features, it is possible to distinguish stylistic signal characteristic of this collection, as well as locate some stylistic heterogeneity within its boundaries.

Methods

The proposed study relies heavily on the methods applied in the context of similar research questions in Camps, Clérice, and Pinche (2021), where a corpus of short and mostly anonymous Old French texts was analysed to uncover original authorial series obscured by layers of compilatory work. Moreover, this study demonstrated the validity of HTR-based data extraction pipelines for stylometric analysis. Below, I discuss the key implementation details of relevant stylometric tests and data preparation.

Extended features: POS *n*-grams and pseudo-affixes

While some stylometric tests proposed in recent literature perform well in authorship attribution tasks for samples much shorter than 3,000 words, these solutions largely rely on word embeddings and training author-specific classifiers.²⁰ These techniques, in turn, require framing the problem as a supervised scenario based on a dataset of securely labelled samples, which is not feasible in this case. Instead, I plan to extend the set of analysed features, aiming to obtain richer representations of samples and thus enhance the performance on shorter *quaestiones*.

A strategy suggested in some recent literature is to incorporate Part-of-Speech (POS) *n*-grams (Chen et al. 2024). Of many possible extended features, the POS three-grams are especially promising as a simple representation of syntactic structures mostly ignored in the bag-of-words approach of tests based solely on word distributions. Incorporating POS three-grams is further facilitated by the availability of efficient morphological taggers for Latin. For the proposed study, I intend to use the LatinPipe (Straka, Straková, and Gamba 2024) or closely related UDPipe 2, both of which provide API access and report very high performance on POS tagging (over 99% accuracy), including on scholastic Latin

¹⁷As noted in the stemma, Ca-Cb, unlike Cc-Cf, are independent from γ . This organisation of the material could be further improved by accounting for differences between H / K and α , as well as distinguishing between sections of ms. C transmitting material found also in β . Unfortunately, some of such classes would score below 3,000 words.

¹⁸Notably, the classes displaying distinct stylistic signals are the shortest ones. The longer classes are also likely to contain portions of stylistically diverse material, but their location requires finer data granularity – ideally at the level of individual *quaestiones*.

¹⁹Codex H / K was destroyed during the Second World War and is known today only through low-quality microfilm reproductions, rendering its palaeographic analysis at best tentative.

²⁰For examples, see the discussion of Multi-Author Writing Style Analysis Task at PAN 2024 – Zangerle et al. (2024).

corpora.²¹ Moreover, following Camps, Clérice, and Pinche (2021), I will extend analysed features by pseudo-affixes, i.e., character three-grams representing each word's boundaries,²² which have been shown to provide valuable stylistic signals (Sapkota et al. 2015).

For each individual feature (most frequent words, POS three-grams, prefixes), the minimal statistically reliable sample length will be assessed implementing the test proposed by Moisl (2011), in which once more I follow Camps, Clérice, and Pinche (2021). Establishing this threshold is the study's primary goal and will condition the later analysis of the data since it determines the exact set of *quaestiones* which can be reliably subjected to cluster analysis.

Data preparation

For the proposed study, I will benefit from access to the machine-readable text of most or all of Langton's *quaestiones*. Nevertheless, I also intend to perform tests on HTR-extracted transcriptions. It can be expected that no significant difference in performance will be observed, with the critical edition being effectively a denoising procedure, although we cannot *a priori* rule out the possibility that editorial interventions left some systematic stylistic trace. The primary goal in experimenting with automated transcriptions is to develop workflows facilitating the inclusion of relevant unedited sources (or manuscript-specific versions of edited material) in further stylometric studies. In the immediate context of Langton's corpus, this would offer great aid in the survey of his vast and mostly unstudied scriptural commentaries.

I intend to test in this study the relatively recent transformer-based HTR solutions (TrOCR), which have been successfully applied to historical material (Ströbel et al. 2022). These architectures rely on a vision transformer for feature extraction and a BERT-type decoder for the translation of visual tokens into characters, offering a few relevant advantages over widely applied solutions based on convolutional neural networks. First, they work exceptionally well with normalised transcriptions, largely facilitating the preparation of ground truth. This comes with a significant advantage in the context of university-based Latin literary production, which features a high density of often idiosyncratic abbreviations. In this case, framing the abbreviation expansion as a downstream task performed on HTR-extracted (semi-)diplomatic transcription is considerably more complex than for vernacular corpora, which generally confer less frequent and more regular abbreviations.²³ Moreover, the reliance on a transformer decoder is likely to result in noise reduction: since the model has a high preference for regular forms, it will likely at least partially normalize orthography, facilitating the later lemmatisation task. Even where the transcription is inaccurate, the produced form can be sufficiently close to ground truth to enable correct assignment of POS tags and prefixes. Consequently, the task-specific accuracy of extracted features is likely to be significantly higher than suggested by the reported Character Error Rate of the model, which can be expected to score around 2%–3%.

²¹For the reported performance, see Straka, Straková, and Gamba (2024), Table 4, especially the performance on Index Thomisticus Treebank. The final article will report performance measured on an annotated sample from Langton's collection.

²²To give an example, word "verbum" would generate pseudo-affixes "_ve," "ver," "bum\$," and "um_."

²³For a relevant example of transcription guidelines framing abbreviation expansion as a downstream task, see Pinche et al. (2023).

For ground truth preparation, I plan to rely on Kraken's blla model for text segmentation.²⁴ While Camps, Clérice, and Pinche (2021) reported low performance for segmentation with Kraken's legacy model (default at the time), initial tests show that currently blla outperforms 'Transkribus' Universal Lines in polygonisation, creating overall more spacious line polygons and capturing relevant abbreviation markers. I will reuse the transcriptions provided by the collection's editors, manually aligning a portion of the material (c. 20 pages), after which I will train a provisional Kraken model and automatically align the transcription for remaining pages.²⁵ While it would be convenient to prepare in this way ground truth for all major codices transmitting Langton's collection, I will prioritize workflow exploration over providing a comprehensive dataset.

Potential results

As noted above, the final results of this study will depend heavily on the exact value of the minimal sample length established in statistical tests. It should be noted that this threshold is calculated for every individual feature and depends on the feature's overall probability in the corpus. Consequently, it will be necessary to balance out the exact set of features and corpus composition, almost certainly resulting in the exclusion of some of the shortest *quaestiones*. Depending on the composition of the final corpus, the study will address three questions:

- Can we discern some distinct clusters among longer *quaestiones*? Such clusters would likely correspond to the activity of different editors, potentially including a cluster of *quaestiones* directly corrected by Langton.
- In general, do short and long versions of one *quaestio* tend to cluster together? Such clusters could indicate cases in which either the original *reportator* developed the longer version or in which the longer version preserved *verbatim* most of the *reportatio*. If no clusters of this type are observed, this would suggest a systematic stylistic difference between *reportationes* and literary *quaestiones* beyond the obvious difference in length.
- Finally, if it will be possible to include most of the short *quaestiones*, can we observe any clusters of *reportationes*? Such clusters could be linked to the activity of individual *reportatores*.

Acknowledgements. I am thankful to Magdalena Bieniak and Wojciech Wciórka for reading an earlier version of this article and sharing their helpful remarks. I would also like to thank Gary Macy for sharing his transcription of ms. Bruges 247.

Author Contributions. Jan Maliszewski — Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing.

Data availability statement. The data and code used in this study are available at: <https://github.com/jtmaliszewski/CHR-2025-It-takes-a-village>. Please note that due to unresolved copyright concerns, the plain text data was masked: all but the top 200 most frequent words were replaced with a "MASKEDTOKEN" placeholder. This allows for full reproduction of the exploratory analysis presented in this article, and the unmasked data was disclosed for peer review. I am currently seeking permission from relevant parties to publish the entire unmasked corpus as part of the final research report. In the meantime, if you are interested in inspecting the unmasked corpus, please contact me at

²⁴Documentation available at https://kraken.re/main/api_docs.html

²⁵Automatic transcription alignment, based on PASSIM script for text reuse detection, was implemented in eScriptorium 0.13. On these projects, see Smith (2012–2023), Kiessling et al. (2019).

j.maliszewski@uw.edu.pl. The stylometric analysis employed in this article was implemented with the *stylo* package for R – Eder, Rybicki, and Kestemont (2016).

Ethical standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding statement. This work was supported by the National Science Centre, Poland, project 2022/45/N/HS1/03747.

Competing interests. The author declares none.

Primary sources.

Manuscripts. Stephen Langton, *Quaestiones theologiae*

- A Avranches, Bibliothèque municipale, 230, ff. 12^{ra}–294^{rb}
- B Arras, Bibliothèque municipale, 965 (394), ff. 70^{ra}–157^{vb}
- C Cambridge, St. John's College Library, C.7 (57), ff. 171^{ra}–352^{rb}
(Ca = C, ff. 171–218; Cb = C, ff. 219–282; Cc = C, ff. 283–306;
Cd = C, ff. 307–322; Ce = C, ff. 323–346; Cf = C, ff. 347–352)
- H Chartres, Bibliothèque municipale, 430, ff. 3^r–73^v
- K Chartres, Bibliothèque municipale, 430, ff. 74^{ra}–154^{vb}
- L Oxford, Bodleian Library, Lyell 42
- R Città del Vaticano, Biblioteca Apostolica Vaticana,
Vat. lat. 4297
- S Paris, Bibliothèque nationale de France, lat. 16385
- V Paris, Bibliothèque nationale de France, lat. 14556

Robert of Courson, Summa

Bruges 247 = Brugge, Hoofdbibliotheek Biekorf (Stadsbibliotheek), 247

Editions.

Hugh of St. Victor, *Sententiae de divinitate*. In Ambrogio Piazzoni, “Ugo di San Vittore auctor delle Sentetiae de divinitate”, *Studi Medievali* 23 (1982), 912–55.

Peter Abelard, *Historia Calamitatum*, ed. Jacques Monfrin, Paris: Vrin, 1959.

Peter Comestor, *Lectures on the Glossed Gospel of John*, ed. and tr. David M. Foley, 2024, The Catholic University of America Press: Washington, DC.

Stephen Langton, *Quaestiones theologiae*, Auctores Britannici Medii Aevi (ABMA):

Vol. I, ed. Riccardo Quinto, Magdalena Bieniak, 2014,
(ABMA 22)

Vol. II, ed. Wojciech Wciórka, *in preparation*

Vol. III.1, ed. Magdalena Bieniak, Wojciech Wciórka, 2021,
(ABMA 36)

Vol. III.2, ed. Magdalena Bieniak, Marcin Trepczyński, Wojciech

Wciórka, 2022, (ABMA 40)

Vol. III.3, ed. Magdalena Bieniak, Andrea Nannini, 2024,
(ABMA 45)

Vol. IV, ed. Magdalena Bieniak, Jan Maliszewski, *in preparation*

Thomas Aquinas, *Summa Theologiae, prima pars* (= *Opera omnia iussu impensaque Leonis XIII P. M. edita*, t. 4–5, Roma 1888–1889). Digitised text by R. Busa, E. Alarcón is available from [Corpus Thomisticum](https://www.corpus-thomisticum.org/).

References

- Bériou, Nicole. 2020. “Orality in its Written Traces: Bilingual Reportations of Sermons in France (Thirteenth Century).” *medieval worlds* 12, 169–84. https://doi.org/10.1553/medievalworlds_no12_2020s169.
- Cafiero, Florian, and Jean-Baptiste Camps. 2019. “Why Molie Most Likely Did Write His Plays.” *Science Advances* 5, no. 11: eaax5489. <https://doi.org/10.1126/sciadv.aax5489>.
- Camps, Jean-Baptiste, and Florian Cafiero. 2013. “Setting Bounds in a Homogeneous Corpus: A Methodological Study Applied to Medieval Literature.” *Revue des Nouvelles Technologies de l'Information MASHS* 2011/2012: Modèles et Apprentissage en Sciences Humaines et Sociales, RNTI-SHS-1: 55–84.
- Camps, Jean-Baptiste, Thibault Clérice, and Ariane Pinche. 2021. Noisy Medieval Data, from Digitized Manuscript to Stylometric Analysis: Evaluating Paul Meyers Hagiographic Hypothesis.” *Digital Scholarship in the Humanities* 36, no. Supplement_2: ii49–ii71. <https://doi.org/10.1093/dllc/fqab033>.
- Chen, Sarah, Patrick Burns, Thomas Bolt, Pramit Chaudhuri, and Joseph Dexter. 2024. “Leveraging Part-of-Speech Tagging for Enhanced Stylometry of Latin Literature.” In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, Association for Computational Linguistics, 251–9. <https://doi.org/10.18653/v1/2024.ml4al-1.24>.
- d'Avray, David. 1985. *The Preaching of the Friars: Sermons Diffused from Paris before 1300*. Oxford: Oxford University Press.
- De Gussem, Jeroen. 2017. “Bernard of Clairvaux and Nicholas of Montiaemy: Tracing the Secretarial Trail with Computational Stylistics.” *Speculum* 92, no. S1: S190–225. <https://doi.org/10.1086/694188>.
- Eder, Maciej. 2013. “Does Size Matter? Authorship Attribution, Small Samples, Big Problem.” *Digital Scholarship in the Humanities* 30, no. 2: 167–82. <https://doi.org/10.1093/dllc/fqt066>.
- Eder, Maciej, Jan Rybicki, and Mike Kestemont. 2016. “Stylometry With R: A Package for Computational Text Analysis.” *The R Journal* 8, no. 1: 107–121. <https://journal.r-project.org/archive/2016/RJ-2016-007/index.html>.
- Foley, David M. 2024. “Introduction.” In *Peter Comestor, Lectures on the Glossed Gospel of John*, edited by David M. Foley. Washington, DC: Catholic University of America Press.
- Gregory, Alys L. 1930. “The Cambridge Manuscript of the Questiones of Stephen Langton.” *The New Scholasticism* 4, no. 2: 165–226.
- Hamesse, Jacqueline. 1997. “La Technique de la Reportation.” In *L'enseignement des Disciplines a Facultes Arts*, edited by Olga Weijers and Louis Holtz 405–21. Turnhout: Brepols.
- Kennedy, V. L. 1947. “The Content of Courson's Summa.” *Mediaeval Studies* 9, no. 1: 81–107. <https://doi.org/10.1484/j.ms.2.306561>.
- Kestemont, Mike, Sara Moens, and Jeroen Deploige. 2013. “Collaborative Authorship in the Twelfth Century: A Stylometric Study of Hildegard of Bingen and Guibert of Gembloux.” *Digital Scholarship in the Humanities* 30, no. 2, 199–224. <https://doi.org/10.1093/dllc/fqt063>.
- Kiessling, Benjamin, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 2019. “Escriptorium: An Open Source Platform for Historical Document Analysis.” In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2: 19. <https://doi.org/10.1109/ICDARW.2019.10032>.
- Moisl, Hermann. 2011. “Finding the Minimum Document Length for Reliable Clustering of Multi-Document Natural Language Corpora.” *Journal of Quantitative Linguistics* 18, no. 1: 23–52. <https://doi.org/10.1080/09296174.2011.533588>.
- Pinche, Ariane, Thibault Clérice, Alix Chagué, Jean-Baptiste Camps, Malamatenia Vlachou-Efstathiou, Matthias Gille Levenson, et al. 2023. “CATMuS-Medieval: Consistent Approaches to Transcribing Manuscripts: A generalized set of guidelines and models for Latin scripts from Middle Ages (8th–16th century).” <https://inria.hal.science/hal-04346939>.
- Quinto, Riccardo. 1994. *Doctor Nominatissimus. Stefano Langton (1228) e la Tradizione Delle Sue Opere*. Münster: Aschendorff.
- Roberts, Phyllis Barzillay. 1968. *Stephanus de Lingua-Tonante: Studies in the Sermons of Stephen Langton*. Toronto: Pontifical Institute of Mediaeval Studies.

- Saccetti, Riccardo.** 2016. *Le Reportationes e la Nascita Dell'insegnamento Teologico, XII-XIII Secolo*. Firenze: L.S. Olschki. <https://doi.org/10.1400/249798>.
- Sapkota, Upendra, Steven Bethard, Manuel Montes, and Thamar Solorio.** 2015. "Not all Character N-Grams are Created Equal: A Study in Authorship Attribution." In *Proceedings of the 2015 Conference of the North American Chapter of the Association For Computational Linguistics: Human Language Technologies*, edited by Rada Mihalcea, Joyce Chai, and Anoop Sarkar, 93–102. Association for Computational Linguistics. <https://doi.org/10.3115/v1/N15-1010>.
- Siri, Francesco.** 2013. "Lectio, Disputatio, Reportatio. Note Su Alcune Pratiche Didattiche Nel XII Secolo e Sulla Loro Trasmissione." In *Medioevo e Filosofia. Per Alfonso Maier* 109–128.
- Smith, David.** 2012–2023. "Passim Project." <https://github.com/dasmiq/passim>.
- Straka, Milan, Jana Straková, and Federica Gamba.** 2024. "ÚFAL LatinPipe at EvaLatin 2024: Morphosyntactic Analysis of Latin." In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (It4hala) @ Irec-Coling-2024*, edited by Rachele Sprugnoli and Marco Passarotti, 207–14. Torino: ELRA / ICCL. <https://aclanthology.org/2024.lt4hala-1.24/>.
- Ströbel, Phillip Benjamin, Simon Clematide, Tobias Hodel, and Martin Volk.** 2022. "Transformer-based HTR for Historical Documents." In: Workshop on Computational Methods in the Humanities 2022, Lausanne, 9 June 2022 – 10 June 2022.
- Weijers, Olga.** 2013. *In Search of the Truth. A History of Disputation Techniques from Antiquity to Early Modern Times*. Turnhout: Brepols.
- Weijers, Olga.** 2015. *A scholar's paradise. Teaching and Debating in Medieval Paris*. Turnhout: Brepols.
- Zangerle, Eva, Maximilian Mayerl, Martin Potthast, and Benno Stein.** 2024. "Overview of the Multi-Author Writing Style Analysis Task at PAN 2024." In *Conference and Labs of the Evaluation Forum*. <https://api.semanticscholar.org/CorpusID:271860835>.