Exploration of Medieval Manuscripts through Keyword Spotting in the MENS Project

Hubert Alisade[1], Diego Calvanese[2,3], Mario Klarer[1,*], Alessandro Mosca[2], Nonyelum Ndefo[2], Bernadette Rangger[1] and Aaron Tratter[1]

[1] *Department of American Studies, University of Innsbruck, Austria*
[2] *Faculty of Engineering, Free University of Bozen-Bolzano, Italy*
[3] *Department of Computing Science, Umeå University, Sweden*

Abstract

In-depth searching for specific content in medieval manuscripts requires labor-intensive, hence time-consuming manual manuscript screening. Using existing IT tools to carry out this task has not been possible, since state-of-the-art keyword spotting lacks the necessary metaknowledge or larger ontology that scholars intuitively apply in their investigations. This problem is being addressed in the "Research Südtirol/Alto Adige" 2019 project "MENS – Medieval Explorations in Neuro-Science (1050–1450): Ontology-Based Keyword Spotting in Manuscript Scans," whose goal is to build a paradigmatic case study for compiling and subsequent screening of large collections of manuscript scans by using AI techniques for natural language processing and data management based on formal ontologies. We report here on the ongoing work and the results achieved so far in the MENS project.

Keywords

ontologies, named entity recognition, keyword spotting, medieval manuscripts, medieval brain anatomy and physiology

1. Introduction

Medieval brain anatomy and physiology was not restricted to medical discourses alone, but on the contrary, shaped key aspects of all major areas of medieval learning, including soul theories in theology, epistemology in philosophy, and the role of imagination and memory in literature. Although major classical and medieval authorities on brain anatomy and physiology have been relatively well documented by scholars of the history of medicine, the legacy of brain concepts in medieval philosophy, theology, and literature—specifically in cross-disciplinary sources, such as *quodlibeta*, *quaestiones disputatae*, and *commentarii*—have received little or no attention so far.

One reason for this is that in-depth searching for specific content in medieval manuscripts, in this case brain anatomical and physiological references, requires labor-intensive, hence

22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023) – Discussion Papers
* Corresponding author.
hubert.alisade@uibk.ac.at (H. Alisade); diego.calvanese@unibz.it (D. Calvanese); mario.klarer@uibk.ac.at (M. Klarer); alessandro.mosca@unibz.it (A. Mosca); bernadette.rangger@uibk.ac.at (B. Rangger); aaron.tratter@uibk.ac.at (A. Tratter)
0000-0003-4243-351X (H. Alisade); 0000-0001-5174-9693 (D. Calvanese); 0000-0003-0712-9328 (M. Klarer); 0000-0003-2323-3344 (A. Mosca); 0000-0002-1634-9835 (A. Tratter)

time-consuming manual manuscript screening. Using existing IT tools to carry out this task has not been possible, since state-of-the-art keyword spotting lacks the necessary metaknowledge or larger ontology that scholars intuitively apply in their investigations. This problem is being addressed in the "Research Südtirol/Alto Adige" 2019 project "MENS – Medieval Explorations in Neuro-Science (1050–1450): Ontology-Based Keyword Spotting in Manuscript Scans,"[1] which is carried out jointly by the Department of American Studies at the University of Innsbruck, Austria (principal investigator: Mario Klarer) and the KRDB Research Centre for Knowledge and Data at the Free University of Bozen-Bolzano, Italy (co-investigator: Diego Calvanese). The goal of the MENS project is to build a paradigmatic case study for the compilation and subsequent screening of large collections of manuscript scans by using AI techniques for natural language processing and data management based on formal ontologies.

Specifically, the MENS project pursues two independent but interconnected goals:

The first goal is to investigate the manifestations and repercussions of medieval brain anatomical and physiological thinking in a large cross-disciplinary sample corpus, including medieval medical texts as well as hitherto neglected medieval nonmedical philosophical, theological, encyclopedic, lexicographical, and literary texts.

The second goal is interlinked with finding, documenting, and transcribing the corpus. It explores knowledge representation and machine-learning technologies—in particular, ontology-based keyword spotting functions—to screen large amounts of manuscript image data for specific content. This should facilitate the search for specific brain anatomical and physiological references and at the same time provide generic tools for algorithmic-based searches in manuscripts. Hence, the sample corpus serves as a paradigmatic case study for any large-scale computer assisted content searches in Latin manuscript corpora in general.

The project is still running, and the activities carried out so far have revealed technical and methodological challenges in the application of the AI technologies mentioned above which we describe in Section 2. In Section 3, we briefly address how the results of the project will be used for philological research.

2. Use of AI for the Exploration of Medieval Manuscripts

The overall philological inquiry of the MENS project rests on a set of interrelated aspects that comprise web-based development to access large manuscript collections available online, recognition of handwritten text, ontology compilation, and ontology-based keyword spotting, all of which contributes to optimize keyword and content spotting in the manuscripts of interest. For this purpose, the project gathered a team of experts in data management and knowledge representation on the one hand and philologists, experts in medieval philosophy and theology, experts in medieval Latin translations from Greek and Arabic, and experts in medieval literature on the other hand. This highly interdisciplinary team of scholars has been collaborating in carrying out the following key steps necessary to optimize the search for specific content in manuscripts:

1. corpus compilation
2. automated downloading and uploading of manuscript scans

[1] https://www.uibk.ac.at/projects/mens/

3. manuscript image analysis and handwritten text recognition
4. keyword spotting
5. named entity recognition
6. use of an ontology of medical terms

We now describe each of the above steps in more detail, pointing out how AI technologies play a key role but also pose challenges that still need to be addressed.

2.1. Corpus Compilation

As larger *in situ* library searches were beyond the scope of the project, the MENS teams focused on manuscripts whose scans are freely available online. The *Codices Palatini latini*[2] are among the most important Latin manuscripts of the Middle Ages and the early modern period, ranging primarily from the 12th to the 16th centuries. They are meant to serve as a paradigmatic case study for a wide array of research approaches in all areas of manuscript scholarship that require screenings of large image data sets. All of the more than 2,000 Latin manuscripts are available online as high-resolution scans. Of these, 260 manuscripts (Cod. Pal. lat. 1079–1339) with more than 100,000 pages have obvious medical content and served as our initial core corpus for optimizing keyword and content spotting.

2.2. Automated Downloading and Uploading of Manuscript Scans

Before we could carry out the image analysis of the manuscript scans and the automated handwritten text recognition (see Subsection 2.3) as well as the subsequent keyword spotting (see Subsection 2.4), we had to manually download the manuscript scans from websites where they are made available for manual browsing and then manually upload them to the Transkribus software. In order to facilitate time-efficient information handling, we have automated this process in a prototype implementation for the sample corpus. When the proper manuscript scans are downloaded, they are also enriched with all pertinent metadata (e.g., information on author, scribe, title of manuscript, script type, provenance, number of pages, date) that is available on the website.

We are currently also exploring the possibility of downloading scans from a wider range of websites by means of suitable web-crawler algorithms. However, a fully automated approach that would work for generic websites appears unfeasible, given the huge individual differences in the websites that host manuscripts. One possibility to ease the manual burden of extending the set of supported websites would be to rely on wrapper generation technology [1, 2].

2.3. Manuscript Image Analysis and Handwritten Text Recognition

The project uses the Transkribus platform[3] [3, 4] for manuscript image analysis, automated Handwritten Text Recognition (HTR), and subsequent Keyword Spotting (KWS). Transkribus was developed in the EU Horizon 2020 project "READ: Recognition and Enrichment of Archival

[2] https://digi.ub.uni-heidelberg.de/en/bpd/virtuelle_bibliothek/codpallat/signatur/1-199.html
[3] https://readcoop.eu/transkribus/

Documents"[4] and initially released as an open-source tool[5]. Since the end of the Horizon 2020 project, the European Cooperative Society (SCE) READ-COOP SCE[6] has been running and further developing the Transkribus platform.

The following features of Transkribus are essential for the MENS project:

1. Linking text and image: The Transkribus platform provides an expert interface for manuscript transcription linked to the scanned image via polygonal chains at line or word level.

2. Layout Analysis: Before transcribing the uploaded manuscripts, the images need to be divided into text regions and lines. Transkribus carries out this step in the automated Layout Analysis[7]. In most cases, the coordinates do not require manual correction.

3. Handwritten Text Recognition (HTR): After the Layout Analysis, documents can be automatically recognized using the HTR tool of Transkribus [3, 4]. The MENS project automatically recognized twenty-one manuscripts of the *Codices Palatini latini*[8] by using the CITlab HTR+[9] model "Gothic_Book_Scripts_XIII-XV_M4,"[10] which has been trained on Latin and German manuscripts from the 13th to the 15th centuries.

4. Tagging: Standardized textual tagging, such as cases of doubt, abbreviations, interpretations, obscurities, personal names, titles, and terminology, is possible and allows for the creation of standardized reference databases.

5. Standards: All data is saved in XML files containing the line coordinates and the transcription including textual tags. From this, various other formats can be generated, e.g., the format of the Text Encoding Initiative (TEI)[11], an internationally acknowledged standard for digital transcriptions and editions.

2.4. Keyword Spotting

The now-defunct Keyword Spotting (KWS)[12] function in Transkribus made it possible to search for words in texts that were automatically recognized using a CITlab HTR+ model. In contrast to full-text search, KWS is able to find the searched-for words even if they are spelled incorrectly in the transcription. This is possible because the tool uses all probability values for each character and not only the most probable result. The recognized text shows only the characters with the highest probability value, but the KWS function considers the values for the second, third, etc. most probable character as well. Besides the searched-for keyword, extracted information for each search result includes the document, the page number, and the line in which the keyword was found, as well as the transcription of the line, and a Confidence Value (CV)[13] between 0

[4] https://cordis.europa.eu/project/id/674943
[5] https://github.com/Transkribus; https://gitlab.com/readcoop
[6] https://readcoop.eu/
[7] https://readcoop.eu/glossary/layout-analysis/
[8] Cod. Pal. lat. 1082, Cod. Pal. lat. 1084, Cod. Pal. lat. 1085, Cod. Pal. lat. 1102, Cod. Pal. lat. 1104, Cod. Pal. lat. 1111, Cod. Pal. lat. 1118, Cod. Pal. lat. 1119, Cod. Pal. lat. 1150, Cod. Pal. lat. 1151, Cod. Pal. lat. 1159, Cod. Pal. lat. 1161, Cod. Pal. lat. 1166, Cod. Pal. lat. 1174, Cod. Pal. lat. 1218, Cod. Pal. lat. 1227, Cod. Pal. lat. 1274, Cod. Pal. lat. 1281, Cod. Pal. lat. 1308, Cod. Pal. lat. 1315, Cod. Pal. lat. 1337
[9] https://readcoop.eu/glossary/htr-plus/
[10] https://readcoop.eu/model/latin-and-german-gothic-book-scripts/
[11] https://tei-c.org/
[12] https://readcoop.eu/glossary/keyword-spotting-kws/
[13] https://readcoop.eu/glossary/confidence-value/

and 1 that represents the accuracy of the tool in finding the searched-for word. The higher the CV is, the higher the probability that the result coincides with the request.

In the twenty-one automatically recognized manuscripts of the *Codices Palatini latini*, the KWS yielded 53,790 hits with a CV $\geq$ 5% for eight searched-for keywords[14] from the domain of brain anatomy and physiology, including some hits not related to the actual text such as later markings and library notes. 3,900 of these hits[15] have a CV $\geq$ 10%, which make up 7.25% of the hits.

The quality and quantity of the hits differs considerably between the individual keywords. This applies not only to the distribution of the CVs but also to the number of false positive hits. The keyword *ymaginacio* has the fewest hits (568) with a CV $\geq$ 5%. On the one hand, this word might occur less frequently in the manuscripts compared to the other keywords. On the other hand, the KWS might have yielded fewer hits because the letter sequence *ymaginacio* is intrinsically more unusual and therefore yields fewer false positive hits. 178 of the 568 hits (31.34%) for *ymaginacio* have a CV $\geq$ 10%. For the keyword *estimatiua*, 418 of the 9,754 hits (4.29%) have a CV $\geq$ 10%, much less in relative terms than for *ymaginacio*. For *sensus communis*, only 33 of the 1,557 hits (2.12%) have a CV $\geq$ 10%. This is the lowest relative proportion of all keywords. The quality of the hits is reflected in the fact that *ymaginacio* still has many true positive hits with a CV $<$ 10%, while *estimatiua* has some false positive hits with a CV $\geq$ 20%. This indicates that the usefulness of the keywords varies. In general, the results show that words with a less frequent letter sequence in the respective language lead to better results. This concerns both the distribution of CVs in favor of higher values and the number of true positive hits.

2.5. Named Entity Recognition

Named Entity Recognition (NER) is a Natural Language Processing (NLP) task that aims at identifying entities of interest for the application domain within unstructured text given as input. Specifically, in the MENS project we have implemented an entity ruler, i.e., a rule-based natural language component that searches for entity names in a Latin text based on a set of predefined rules (or patterns) and assigns to the identified entities a corresponding descriptive tag. In defining the entity ruler, eight distinct types of pattern groups have been specified to discover named entities: persons, groups, places, diseases, body parts, status, senses, and brain. The first three patterns, specified for extracting demographic information, have been compiled from several resources on the Web. The remaining patterns relating to anatomy and physiology have been defined based on the information compiled in a specific thesaurus, itself updated with data compiled from other thesauri on the Web. Since Latin is an inflected language involving declensions, creating these patterns also inspired the need for an automated decliner to improve the discovery of named entities.

One of the limitations of the entity ruler was its rigidness, i.e., it could only recognize the specified patterns but nothing more. To address this issue, we used a custom machine learning-

[14] *cerebrum* (10,012 hits), *estimatiua* (9,754 hits), *fantasia* (11,789 hits), *memoria* (13,409 hits), *sensus communis* (1,557 hits), *spiritus animalis* (2,359 hits), *uentriculum* (4,342 hits), *ymaginacio* (568 hits)

[15] *cerebrum* (1,369 hits), *estimatiua* (418 hits), *fantasia* (519 hits), *memoria* (1,148 hits), *sensus communis* (33 hits), *spiritus animalis* (72 hits), *uentriculum* (163 hits), *ymaginacio* (178 hits)

based solution that can intelligently identify named entities in data that it has never seen before. We needed to create a training data set using the entity ruler and data extracted from two libraries: the *Perseus Digital Library*[16], specifically texts from the collection [5], and a selection of books from *The Latin Library*[17]. These data sets were used specifically because they focus on medical and basic demographic data. The training data, loaded with examples of named entities, was then used to create an NER model. The result of this task is the application of the NER model on unseen data with the expectation that it will (correctly) recognize entities in a given text.

2.6. Use of an Ontology of Medical Terms

To further improve the keyword spotting function for medical content in medieval Latin manuscripts, we started the design of a domain ontology including pertinent medieval anatomical, physiological, and further medical terminology on the basis of the 3,718 entries in the specific anatomical glossary by Adolf Fonahn [6], supported by the extremely valuable but still incomplete online *Arabic and Latin Glossary*[18] edited by Dag Nikolaus Hasse, as well as the 13th-century *Clavis sanationis*[19] by Simon of Genoa. The expected resulting ontology will include knowledge that uninitiated researchers might overlook when using keywords in a simple lemma-based search. For example, in medieval medicine, at least in Aristotelian circles, mental processes were thought not to originate from the brain alone but also and even in principle from the heart. The ontology will take into consideration these kinds of content-based aspects of possible searches. In order to make the ontology as functional as possible, besides the medieval brain anatomical and physiological concepts, it will be extended to also include a considerable number of other important medieval medical terms. With its larger general scope, the structure of the ontology becomes a paradigmatic tool for research in medieval medicine that is applicable to other medical corpora beyond the brain anatomical and physiological focus of the project.

When a manuscript is examined for multiple keywords of a domain, the hits can be used to check whether the manuscript contains content related to that domain. This works better the lower the rate of false positive hits is and the more specific the keywords are. This needs to be taken into account when using an ontology to improve the keyword spotting function. Homographs are therefore rather unsuitable, especially if the same spelled words occur frequently in common parlance. Personal and place names that do not otherwise occur in the language are likely to be particularly suitable for searching. Therefore, historical documents could be searched for them in this way.

Having an ontology is also a first step toward the application of the Ontology-Based Data Access (OBDA) paradigm [7, 8] to ease the access to the information contained in the recognized manuscripts. In OBDA, users are provided with a domain ontology expressed in a lightweight ontology language [9], which is then exploited to expand the initial query in a semantically consistent way. If one looks, e.g., for instances of concept A, and A is declared in the ontology

[16] https://www.perseus.tufts.edu/hopper/
[17] https://www.thelatinlibrary.com/
[18] https://algloss.de.dariah.eu/
[19] http://www.simonofgenoa.org/index.php?title=Simon_Online

as having synonym B and subclass C, the system will also be able to automatically return instances of B and C when asked for A. Similarly, in presence of domain-related relationships connecting concepts in the ontology, such as 'metaphor for,' 'responsible for,' or 'located in,' users will be able to specify queries by explicitly using them (e.g., "Show me all the occurrences of X that are ⟨responsible for⟩ the ⟨imagination⟩") and rely on the system to retrieve parts of the text whose syntactic shape is initially unknown but explicitly specified in the ontology.

3. Use of the Corpus of Source Texts

The legacy of the project will be guaranteed by the integration of all pertinent findings of the keyword searches into a compilation and transcription (with tags and metadata) of a brain anatomical and physiological corpus of scientific source texts—a corpus that fulfills a number of interrelated but distinct functions:

- Compilation of an annotated bibliography or *repertorium* of pertinent brain anatomical and physiological sources in different scholarly disciplines (medicine, philosophy, theology, encyclopedia, literature, etc.) and five different languages (Greek, Latin, Arabic, Hebrew, and Syriac) in the Middle Ages.
- Through the implementation of proof-read machine transcriptions of brain relevant passages from manuscripts and already existing editions, the *repertorium* will also fulfill the function of an anthology or computer-searchable corpus of source texts.
- Standardized tagging and metadata accumulation augments the corpus of transcribed texts into a reference tool of medieval brain knowledge. This will include names of authorities, straightforward medical terms, and metaphorical uses to denote and describe certain faculties of the human brain (e.g., in the Latin version of Avicenna's *Canon of Medicine* [10] the expression *nuntius et vicarius* ("messenger and deputy") is used for the spinal cord).
- Transcriptions, tags, and concordances of brain anatomical and physiological terminology will allow us to trace lines of influence. Since many texts do not mention their sources explicitly, influences can only be reconstructed by comparing the uses of terminologies or metaphors (as in the above example from *The Canon of Medicine*).

4. Conclusion

In this paper, we reported on the ongoing work and the results achieved so far in the MENS project. Using AI technologies for handwritten text recognition, keyword spotting, and named entity recognition shows promising results in finding specific content in medieval manuscripts, in this case brain anatomical and physiological references. In contrast to full-text search in automatically recognized medieval manuscripts, the keyword spotting function has a much higher success rate in finding searched-for words and therefore has significant advantages over other technologies in spotting specific content. As handwritten text recognition is rapidly improving, keyword spotting results are becoming more precise, making it easier to find the content of interest. A corpus of tagged transcriptions along with an ontology of medical terms will serve as a reference tool for medieval brain knowledge, allowing lines of influence to be traced across the Middle Ages.

Acknowledgments

The project "MENS – Medieval Explorations in Neuro-Science (1050–1450): Ontology-Based Keyword Spotting in Manuscript Scans" was funded by the Autonomous Province of Bolzano/Bozen – Department Innovation, Research, University and Museums under the research program "Research Südtirol/Alto Adige" 2019 (funding contract number 14/34). Diego Calvanese has also been partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

[1] R. Baumgartner, S. Flesca, G. Gottlob, Supervised Wrapper Generation with Lixto, in: Proceedings of the 27th International Conference on Very Large Data Bases (VLDB), Morgan Kaufmann Publishers, 2001, pp. 715–716.

[2] M. Bronzi, V. Crescenzi, P. Merialdo, P. Papotti, Wrapper Generation for Overlapping Web Sources, in: Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, volume 1, IEEE, 2011, pp. 32–35. doi:10.1109/WI-IAT.2011.160.

[3] P. Kahle, S. Colutto, G. Hackl, G. Mühlberger, Transkribus – A Service Platform for Transcription, Recognition and Retrieval of Historical Documents, in: Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), volume 4, IEEE, 2017, pp. 19–24. doi:10.1109/ICDAR.2017.307.

[4] G. Mühlberger, L. Seaward, M. Terras, et al., Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study, Journal of Documentation 75 (2019) 954–976. doi:10.1108/JD-07-2018-0114.

[5] Celsus, On Medicine, translated by W. G. Spencer, Harvard University Press, Cambridge, MA, 1935–1938. 3 volumes, Loeb Classical Library 292, 304, 336.

[6] A. Fonahn, Arabic and Latin Anatomical Terminology: Chiefly from the Middle Ages, Jacob Dybwad, Kristiania, 1922.

[7] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, R. Rosati, Linking Data to Ontologies, in: S. Spaccapietra (Ed.), Journal on Data Semantics X, volume 4900 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 2008, pp. 133–173. doi:10.1007/978-3-540-77688-8_5.

[8] G. Xiao, D. Calvanese, R. Kontchakov, D. Lembo, A. Poggi, R. Rosati, M. Zakharyaschev, Ontology-Based Data Access: A Survey, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI-18), International Joint Conferences on Artificial Intelligence Organization, 2018, pp. 5511–5519. doi:10.24963/ijcai.2018/777.

[9] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, R. Rosati, Tractable Reasoning and Efficient Query Answering in Description Logics: The *DL-Lite* Family, Journal of Automated Reasoning 39 (2007) 385–429. doi:10.1007/s10817-007-9078-x.

[10] Avicenna, Liber canonis Avicenne revisus et ab omni errore mendaque purgatus summaque cum diligentia impressus, translated by Gerard of Cremona, Paganino Paganini, Venice, 1507.