Studying co-occurrences of texts in Middle Dutch prayer books

Peter Verhaar[1], Anna Dlabačová[1], and Susanne de Jong[1]

[1] Centre for the Arts in Society, Leiden University, Leiden, The Netherlands

Abstract

This study explores the shared transmission of texts and the formation of possible groups of Middle Dutch prayer texts in Books of Hours. Using computational methods such as Pointwise Mutual Information, Jaccard similarity, and Levenshtein edit distance, the study identified statistically significant groupings and clusters of co-transmitted texts. These analyses revealed recurring structures, degrees of similarity between books, and regional compilation patterns. This study of the shared transmission of different books helps to shed light on devotional practices and reading behaviour in the late Middle ages.

Keywords: Prayer books, Devotion, PMI, Edit distance, Textual co-transmission

1 Introduction

Prayer books were by far the most commonly read Dutch-language books in the late medieval Low Countries, and the most popular prayer books were undoubtedly Books of Hours, in the Middle Dutch translation ascribed to Geert Grote, the founder of the *Devotio Moderna*. Surviving in ca. 850 manuscripts and ca. 40 printed editions,[1] Grote's translation must have had a profound impact on devotional culture in the Middle Ages. Despite this clear proliferation in the number of books, the scope and the abundance of vernacular prayer books has scarcely been investigated by scholars focusing on Middle Dutch literature. The ERC-funded project *Pages of Prayer* is currently carrying out the first large-scale investigation of this unique corpus of Dutch-language prayer books.

Studying Middle Dutch prayer books is challenging at present, because the information about the extant books is scattered across many different siloed resources. One important resource is the inventory of Middle Dutch manuscripts held at the Royal Library in Brussels, which catalogues over 300 prayer texts [5]. A second important source is the *Bibliotheca Neerlandica Manuscripta* (BNM) which was set up as a paper database at the beginning of the twentieth century by the Flemish philologist and palaeographer Willem De Vreese (1869-1938). The BNM was converted into a digital database in the 1990s, and, after 2013, the BNM was expanded into the BNM-I.[2] The BNM-I database focuses principally on texts, but it also contains basic codicological information on the production and ownership of manuscripts. Relevant – mainly bibliographical – information related specifically to printed prayer books can be found in large online bibliographies such as the

Peter Verhaar, Anna Dlabačová, and Susanne de Jong. "Studying co-occurrences of texts in Middle Dutch prayer books." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 1564–1570. https://doi.org/10.63744/TsDYFPy1wlfMm.

* This research was carried out as part of the project 'Pages of prayer: The ecosystem of vernacular prayer books in the late medieval Low Countries, c. 1380–1550' (2023–2028) supported by ERC grant PRAYER, Grant agreement no. 101041517, funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

[1] Estimates differ, from 800 to over 2000 manuscripts [3, p. 161].

[2] The intention was originally to include data on printed books as well, but this has not materialised so far.

ISTC (*Incunabula Short Title Catalogue*), the STCN (*Short Title Catalogue Netherlands*), USTC (*Universal Short Title Catalogue*) and the GW (*Gesamtkatalog der Wiegendrucke*).

During a first phase of the *Pages of Prayer* project, this existing information about prayer books has been aggregated and integrated. In the database that has been created for this project, the various descriptions have been harmonised and expanded, in many cases quite drastically. The data have been structured according to a custom-made data model [4] that was implemented in *Heurist*, an online database management system. Important entities include 'Book', 'Production Layer', 'Text' and 'Expression'. One 'Book' – that is, the object as it has survived – can contain one or more 'Production Layers' (handwritten or printed), which can in turn contain witnesses of texts. These witnesses are referred to as 'Expressions'. For each 'Expression', data is collected about the incipit, the rubrics, the presence of abbreviations and glosses in the text, among many other aspects. Each 'Expression' is linked to an authority record named the 'Text'. This latter entity describes the text in its abstracted typical form. A crucial property of the 'Text' is the identifier, and the standardised title. The identification numbers have been copied from the Brussels inventory of Middle Dutch manuscripts. For rhymed prayers, we have adopted the identification numbers minted by Johan Oosterman in his study of Middle Dutch rhymed prayers in manuscripts [6]. Texts lacking an existing identifier have been assigned new numbers, following the 'P' prefix.

One of the main aims of the project is to conduct a longitudinal analysis of Middle Dutch prayer texts transmitted via different kinds of media, and to chart patterns that can tell us more about the preferences and the developing religious practices of their readers. One of the subquestions, central to this goal, concentrates on the shared transmission of texts and the formation of possible groups of texts in Books of Hours, which are all compilations containing multiple texts. As the contexts in which these texts were transmitted and used are of crucial importance, this paper aims to examine whether it is possible to identify patterns in the co-occurrence of specific texts in different books. The analysis of such patterns can shed light on devotional preferences and reading behaviour in the late Middle Ages. The co-occurrence of the various Hours and the 'core texts' of a Book of Hours (calendar, Penitential Psalms, Litany and Vigils) in Grote's translation has been studied in the early 1990s by Rudolf van Dijk. He examined the co-occurrences as well as the order in which these texts appear in 171 manuscripts, disregarding all other texts that were transmitted together with Grote's Hours.[3] This paper, by contrast, aims to provide a preliminary examination of the additional or accessory texts (primarily 'loose' prayers).

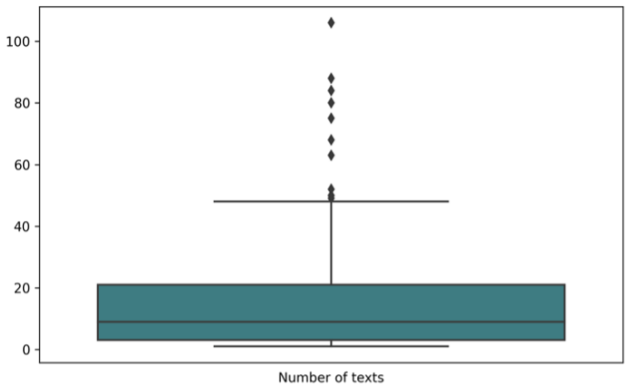

Figure 1: Distribution of the number of texts per book

The PRAYER database currently describes 500 books with handwritten texts and 124 printed books. These books include a total number of 6584 expressions of 1567 texts. The study that is

presented in this paper concentrates on manuscripts only and, as mentioned above, the emphasis is on the texts that were disseminated alongside the Hours (Hours of the Virgin, Hours of the Holy Spirit, Hours of the Holy Cross (long and short), and Hours of the Eternal Widsom) and the 'core texts'. For this reason, the latter texts have been removed from the dataset. Next to this, a number of texts in the dataset have been merged. The texts with the identifiers G004, G004a, G004b, G004c, G004d and G004e are all variants of the "Prayers of St. Gregory to the Arma Christi". This is a prayer that 'grew' from five to ten verses over the course of the fifteenth century. The differences between the versions are highly relevant for the dating of books, but less so for the study of co-occurrences and textual patterns. Identifiers of texts which are identical except for their prologues have been merged as well. These filtering operations resulted in a dataset describing 272 manuscripts, containing 4085 expressions of 1459 texts. Figure 1 gives an impression of the distribution of the number of expressions per book. The books in the dataset contain 15 expressions of 'accessory' or 'additional' texts on average. The Brussels manuscript with shelfmark 12079 and the manuscript from the collection of the Victoria and Albert Museum (London), MSL/1902/1672 (Reid 35) clearly form outliers, with 106 and 88 expressions of additional texts respectively.

2 Groupings

In our conceptualisation of textual co-transmission, we have made a distinction between clusters and groupings. Clusters, firstly, are viewed as fixed combinations of texts occurring in multiple books. A grouping, on the other hand, is a looser form of co-transmission in which two or more texts appear together in multiple books, but not necessarily in the same order. Groupings of texts co-occurring in different books have firstly been identified using *Pointwise Mutual Information* (PMI), a widely used method in NLP to identify related words.[3] In essence, PMI measures whether two events occur together more frequently than would be expected if they were statistically independent. To calculate the PMI value, the probability of the cooccurrence of the two events firstly needs to be divided by the product of the individual probabilities of these events. The final PMI value is obtained by calculating the log2 of this division. If the PMI value is positive, this indicates that the probability of the cooccurrence of the two events is higher than the probabilities that can be expected if these were purely based on chance. For the purpose of this study, PMI values were calculated for each combination of texts. Two texts were taken to co-occur if they are contained within the same book. The probabilities of the individual texts were calculated by dividing the frequencies of the texts (i.e. the number of witnesses) by the total number of expressions in the dataset.[4]

While PMI appears to be effective for retrieving statistically significant cooccurrences of two texts, the calculated values did not directly help us to retrieve larger groupings. To analyse such compounds consisting of multiple texts, we additionally identified sets of texts that can be assumed to be strongly associated. More specifically, we selected, for each text, all the other texts whose associated PMI scores were positive. Within such sets of texts, all possible combinations of these texts were generated, in groupings of lengths ranging from four to twelve texts. These combinations were produced using the *combinations* module from the *itertools* package in Python. It was assumed that groupings of less than four texts are not meaningful. From all the possible groupings, we only retrieved those groupings with a frequency of three or more.

Many of the clusters that were identified using this method corroborate existing expectations about reappearing text groupings within the accessory texts. It was found, for example, that the

[3] The method that was adopted in this study to identify reappearing grouping largely follows the approach that is discussed in Gleb Schmidt, "Computational Approaches to the Study of Patristic Sermon Collections" [7] and in David Birnbaum, "Computer-Assisted Analysis and Study of the Structure of Mixed Content Miscellanies" [1].

[4] The code that was developed for this study can be found at `https://github.com/peterverhaar/prayer_clusters`

prayers for the deceased often co-occur. The prayers for a deceased Priest or Bishop, a deceased Man, Woman, Person, and Benefactors (G161, G162, G163, G164, G166) are copied jointly in 6 books. The texts with identifiers G016, G013a, G013b, G013c, G017 and G018 – all prayers connected to the Communion and translated from Latin – likewise occur together, in nine books. Figure 2 visualises the books containing this specific grouping and the locations of the texts within these books. The texts that make up the grouping are highlighted in red. It can be seen, among other things, that G013a, G013b and G013c are transmitted in the exact same order in all nine books. These three different prayers can occur separately, but they often form a single entity. The same is true for G017 and G018. The texts of the grouping are transmitted in the exact same order in seven of these books. In one manuscript, another prayer is added in between G013c and G017. In another manuscript, G016 is placed near the end of the book and is followed by G013a, G013b and G013c, which have been copied twice in this manuscript. The marked consistency in the order of these prayers can likely be explained through their connection with the liturgy of the Mass, following the fixed elements of the ritual. Although these results may not be surprising, they remain relevant: the clusters of different communion prayers revealed by the analysis warrant closer textual study to understand how such texts enabled participation in a Latin church ritual through the vernacular.

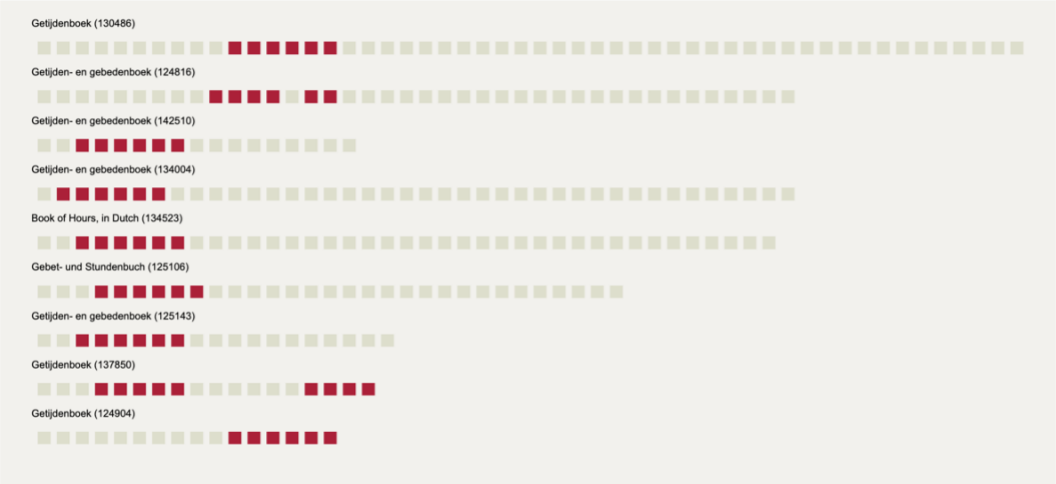

Figure 2: Grouping consisting of G016, G013a, G013b, G013c, G017, G018

The method that was developed also resulted in groupings whose texts do not seem to be related in terms of content. For example, the texts “G004: Prayer of St. Gregory to the Arma Christi”, “G048: Prayer to be recited after the H. Communion or after Mass”, “G105: Prayer to Mary in the Sun with Indulgence Pope Sixtus IV”, “G019: Prayer to be recited before the H. Communion” and “G072: Prayer to be recited before the H. Communion” are jointly included in 5 books. Figure 3 shows that these prayers can be found in different places in each manuscript. In two manuscripts, the three communion prayers occur after each other, and in the same order: G072, G019 and G048. Interestingly, both manuscripts are produced in the same place, namely, Enkhuizen. In the other manuscripts, the communion prayers all appear in a ‘thematic’ group of other communion prayers. The “Prayer to Mary in the Sun” and the “Prayer of St. Gregory” appear in different places, with a notable distance in between them.

3 Clusters

The term ‘grouping’ is, in a sense, a hypernym for the term ‘cluster’. As discussed, the latter term is used to refer to a specific type of grouping in which multiple texts are disseminated in the exact same order and adjacent to each other. As borne out by one of the cases highlighted above, the

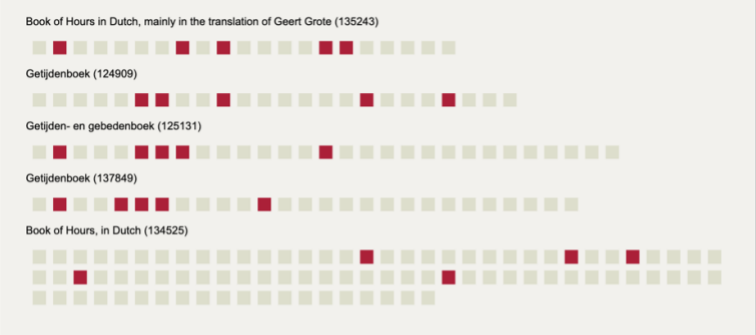

Figure 3: Grouping consisting of G004, G048, G105, G019, G072.

method based on PMI can occasionally uncover clusters of this nature. To be able to retrieve such clusters in a more methodical manner, several experiments have also been conducted, which were inspired, to a large degree, by the principle of extracting ngrams from textual data. The first step in the process was to establish all possible subsequences of expressions, with lengths ranging from three to the total number of expressions in each book. Next, the frequencies of all these fixed order sequences were calculated. Clusters with a frequency of two or less were discarded. When one clusters was found to be a subset of another cluster, the cluster with the lowest frequency was removed from the result set. Figure 4 visualises one of the clusters that could be established using this method. The communion prayers with identifiers G111, G212, and G229 were disseminated as a stable unit in seven manuscripts. While this cluster also includes communion prayers, they differ entirely from those in the previous group, illustrating the diversity and lack of standardisation among the prayers that enabled vernacular participation in a Latin ritual.

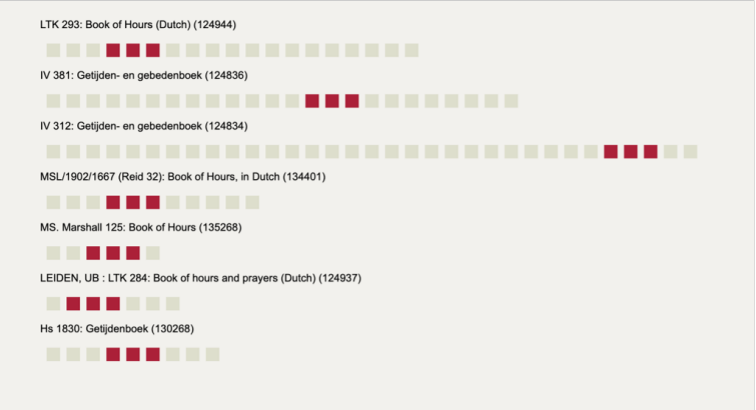

Figure 4: Cluster consisting of G111, G212 and G229

4 Similaritioes between books

Beyond exploring patterns in the co-transmission of texts, the data about the additional prayers texts contained in the prayer books also permit an analysis of the similarities between the manuscripts in their entirety. Among other techniques, the text compilations in the various manuscripts can be compared on the basis of Jaccard similarity. This metric is calculated by dividing the size of the intersection of two collections by the size of their union. A value of 1 indicates that the two collections are identical. To compare the sequences of additional prayers in the dataset, a matrix was created describing the Jaccard similarities for all combinations of two books. The similarities

were subsequently visualised in the form of a network. The nodes in this network were connected if they have a Jaccard similarity score of 0.2 or higher.

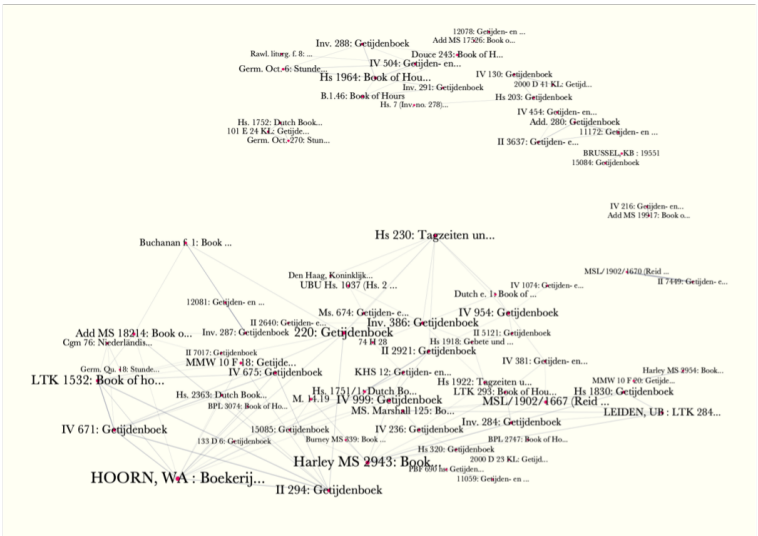

Figure 5: Network displaying Jaccard similarity scores for all manuscripts

As can be derived from figure 5, the Brussels manuscript with shelfmark IV 671 is connected to the Leiden manuscript with shelfmark LTK 1532, having a Jaccard similarity score of 0.6. The six accessory texts in LTK 1532 are all included in IV 671, which additionally contains four other texts. The manuscripts Brussels, KBR, 134004: II 7449 and London, V&A, MSL/1902/1670 (Reid 34) have a Jaccard similarity score of 0.92. They share no less than 38 additional prayers.

Jaccard similarities appear to be effective for the analysis of the level of overlap between two text collections, but it does not take account the order of the texts. A second method that can be used to compare text collections is edit distance. The method was originally developed to quantify the difference between words. The edit distance represents the minimum number of operations that are required to change one word into a second word. In the case of Levenshtein distance, these operations entail the removal, insertion, or substitution of a character in the first string. In the current study, this method has been adopted to assess the similarities between two sequences of texts. The scores have been calculated using the *seqsim* package in Python.[5] Figure 6 displays the books that have an edit distance of 0.4 or less. It was found that the manuscripts Leiden, UL, LTK 1532 and Brussels, KBR, IV 671 share 6 texts. The edit distance is 0.4. These books both contain prayers to be recited before or after the Holy Communion. The manuscripts Hoorn, West-Frisian Archive, Boekerij der gemeente I 45 Oct. and Brussels, KBR, II 294, for example, have a Levenshtein distance of 0.44. They share five texts, and, in the two books, the texts are included in the exact same order.

This paper has discussed various methods for the analysis of co-transmitted texts. At the moment of writing, no definitive conclusions have been drawn regarding the effectiveness of these different approaches. The usefulness of these methods depends on the research context. Methods to identify text clusters can aim to retrieve as many clusters as possible or to selectively identify clusters with specific statistical properties. The preliminary results presented here may, for example, identify clusters that deserve further analysis (such as the diversity within communion prayers) and manuscripts that may or should be studied in relation to each other. Follow-up research may focus on the historical development of specific texts groupings, on the regional preferences for particular compilation practices and on the motivations underlying certain text combinations. It is

[5] https://pypi.org/project/seqsim/

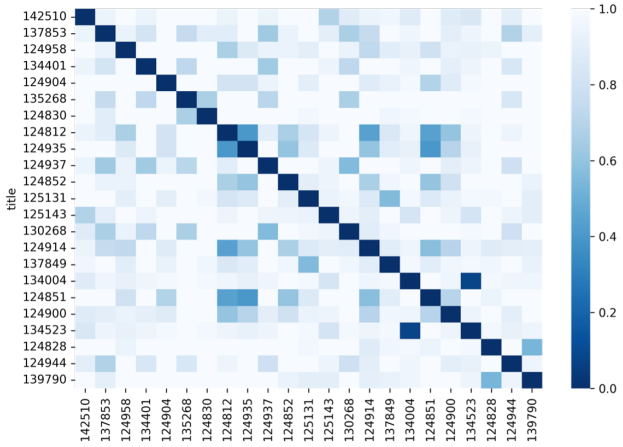

Figure 6: Heatmap displaying Levenshtein distances

hoped that the experimental computational methods, applied initially in an aleatory fashion, can ultimately help to offer a fresh take on the study of the Middle Dutch prayer book.[6].

References

[1] Birnbaum, David. “Computer-Assisted Analysis and Study of the Structure of Mixed Content Miscellanies”. In: *Scripta and E-Scripta 1* (2003), pp. 15–64.

[2] Brantley, Jessica. “How to Look at Lots of Books of Hours: Big Data and Disciplinary Difference”. In: *Digital Philology: A Journal of Medieval Cultures 14* (1 2023), pp. 80–97.

[3] Dijk, R. Th. M. van. “Methodologische kanttekeningen bij het onderzoek van getijdenboeken”. In: *Boeken voor de eeuwigheid. Middelnederlands geestelijk proza*, ed. by Th. Mertens et al. Vol. 2. Amsterdam: Prometheus, 1993, pp. 210–229.

[4] Dlabacova, A. and Eldere, I. van. “Modelling the Middle Dutch Prayer Book (c. 1380–1550): The PRAYER Data Model for Manuscripts and Early Printed Books”. In: *Quaerendo 55* (2025), pp. 43–752.

[5] Herman Mulder, Jan Deschamps en. *Inventaris van de Middelnederlandse handschriften van de Koninklijke Bibliotheek van België*. Vol. 15 vols. Brussels: Koninklijke Bibliotheek van België, 1998.

[6] Oosterman, J. B. *De gratie van het gebed. Overlevering en functie van Middelnederlandse berijmde gebeden*. Prometheus, 1995.

[7] Schmidt, G. “Computational Approaches to the Study of Patristic Sermon Collections”. In: *Augustiniana 73* (2 2023), pp. 43–752.

[6] Examples of such novel approaches are discussed in Brantley, ”How to Look at Lots of Books of Hours: Big Data and Disciplinary Difference” [2]