The Journal of
HISTORICAL
NETWORK
RESEARCH

9 | 2023
special issue

Networks of Manuscripts, Networks of Texts

EVINA STEIN | GUSTAVO FERNÁNDEZ RIVA

C²DH
LUXEMBOURG CENTRE FOR
CONTEMPORARY AND DIGITAL HISTORY

UNIVERSITY OF
LUXEMBOURG

Imprint

Université du Luxembourg 2023

Luxembourg Centre for Contemporary and Digital History (C²DH)
Université du Luxembourg
Belval Campus
Maison des Sciences Humaines
11, Porte des Sciences
L-4366 Esch-sur-Alzette

Editors
Asst. Prof. Dr. Marten Düring (Luxembourg Centre for Contemporary and Digital History | C²DH)
PD Dr. Christian Rollinger (Universität Trier)
Dr. Cinderella Petz (IEG – Leibniz-Institut für Europäische Geschichte)
Dr. Ingeborg van Vugt (Königlich Niederländische Akademie der Wissenschaften)
Clemens Beck, M. A. (Friedrich-Schiller-Universität Jena)

ISSN 2535-8863

Contact
Principal Contact
JHNR-editors@historicalnetworkresearch.org
Support Contact
Dr. Marten Düring (Université du Luxembourg)
JHNR-support@historicalnetworkresearch.org

Cover Design and Typesetting
text plus form, Dresden, Germany
Cover image
Evina Stein
Copyediting
Andy Redwood, Barcelona, Spain

Published online at
https://doi.org/10.25517/jhnr.v9i1

The Journal of HISTORICAL NETWORK RESEARCH

INA SERIF

From Networks of Texts to Networks of Topics?

On the Classification of (Texts in) Compilations with a View towards Manuscript Transmission

Journal of Historical Network Research 9 (2023) 184–213

Keywords manuscripts, topic modeling, shared manuscript transmission, digital history, medieval studies, Jakob Twinger von Königshofen

Abstract As medieval manuscripts often consist of more than one text, the application of network analysis can show textual connections between codices and therefore shed light on the circulation of texts, of manuscripts, and thus of knowledge. However, a text-based analysis often faces difficulties resulting from insufficient manuscript descriptions and a lack of normalization of work titles. A broader view, which would compare not particular texts, but rather genres, areas of interest or fields of knowledge, may help to circumvent these problems; however, this broader approach must deal with problems regarding classification. Instead of finding connections between subjectively classified texts, one can make use of topic modeling as a means to computationally classify, and thus characterize, multiple-text manuscripts. On the basis of automatically detected topics, topic-based networks can be generated. The current potential of such an analysis was tested using a sample of codices that contain the late medieval chronicle of Jakob Twinger von Königshofen. Advancements in text recognition and normalization of non-standardized spelling could further enhance this method to investigate the connections between the codices of a specific corpus and develop a better understanding of the copying and transmission of premodern manuscripts.

C²DH LUXEMBOURG CENTRE FOR CONTEMPORARY AND DIGITAL HISTORY

UNIVERSITY OF LUXEMBOURG

I. Introduction

Researching the textual transmission of a medieval work is often arduous, as it requires dealing with different textual manifestations.[1] A stemmatological approach tries to follow copying processes in order to connect textual witnesses to each other and classify them in relation to an (often imagined or constructed) original.[2] But such a text-based approach has its limits. Texts that could be classified as manifestations of a work are not always identified as such, for various reasons: existing manuscript descriptions may qualify it as unique, due to a lack of knowledge of the describer, and provide it with a new title (or no title at all); modifications during the copying process, like dialectal adaptations, may have altered a text so significantly that connections to related versions are blurred; the layout may also fail to signal the start of a new textual unit, making it hard to identify a certain segment of a manuscript as the manifestation of a specific work. Even if we had the full, searchable text for all handwritten codices available, a textual comparison would be extremely difficult due to the variance of medieval manuscript culture.[3] Considering that the majority of medieval manu-

Acknowledgements: Apart from the credit that goes to Evina Stein and Gustavo Riva for organizing the conference and planning this volume, special thanks go to Gustavo Riva for discussing my sometimes rather unstructured thoughts at different stages of writing this paper. I would also like to thank the two anonymous reviewers for their close reading of the paper and their very helpful comments.

Corresponding author: Ina Serif, ina.serif@unibas.ch

1 In the absence of a conclusive definition of the term 'work', I refer to it as a virtual denominator for a concrete textual representation, but without the intention to value the single text as 'better' or 'closer' to an imagined 'original' in the tradition of Karl Lachmann. For a discussion of the terms 'work' and 'text' and a convincing way out of the Lachmannian idealization of the work or 'Urtext', see Tjamke Snijders, "Work, Version, Text and Scriptum: High Medieval Manuscript Terminology in the Aftermath of the New Philology," *Digital Philology. A Journal of Medieval Cultures* 2, no. 2 (2013): 266–96. She modifies the term 'scriptum', coined by John Dagenais, making it a description of a "material unity of text, layout, and codicology" that can be related to others: "[...] it becomes possible (though not always necessary) to judge them [scripta, I. S.] as relatively similar to one another on a textual level (as variants or even attempted copies) or as more profoundly or characteristically different from one another on the textual plane (versions)." Ibid., 279–280, 285.

2 An attempt to connect manuscripts without necessarily linking them to a root (an original) after normalizing dialectal variance in the transmission of the *Parzival* can be found in Michael Stolz, "Linking the Variance. Unrooted Trees and Networks," in *The Evolution of Texts. Confronting Stemmatological and Genetical Methods. Proceedings of the International Workshop Held in Louvain-La Neuve on September 1–2, 2004,* ed. Caroline Macé, Linguistica Computazionale 24/25 (Pisa, 2006), 193–213.

3 See Bernard Cerquiglini, *Éloge de la variante. Histoire critique de la Philologie* (Paris, 1989); Stephen G. Nichols, "What is a Manuscript Culture? Technologies of the Manuscript Matrix," in *The Medieval Manuscript Book: Cultural Approaches,* ed. Michael Johnston and Michael Van Dussen, Cambridge Studies in Medieval Literature 94 (Cambridge, 2015), 34–59.

scripts contain more than one text,[4] one could try to balance this shortcoming by tracing shared transmissions. By following the copying processes of not just one work but several, or finding connections between 'scripta'[5] (and therefore connections between manuscripts through more than one text), one could for instance discover previously unknown textual testimonies for the researched work in question – this would mean more texts that are used for comparison, which could compensate for imperfect or incomplete manuscript descriptions.[6] Particularly short texts that are not clearly separated in a manuscript copy, but that occur together with other texts, could be traced more easily. However, this approach is also constrained, simply because of temporal capacities.[7] So, instead of relying on attributed work titles, a widening of the perspective on a codex and its contents could make up for inconsistencies or incompleteness. Such a broadened perspective would classify a codex, or rather its content, with additional properties, making it comparable with others on a non-textual level by adding additional metadata, such as the genre(s) of the containing texts. This would enable the exploration of another level of connection between manuscripts, and for the creation of networks between them based on this new metadata. In the following, different ways of codex classification and their advantages and shortcomings will be discussed, using the manuscript transmission of the late medieval German chronicle by Jakob Twinger von Königshofen. I will discuss human created classifiers as well as computational classifiers, each with a view towards the potentials of network analyses based on the specific classifiers. First, networks based on connections between texts will be examined, followed by an analysis of networks between genres. The application of topic modeling will then be proposed as a starting point for a topic-based network, discussing it as one option to detect manuscript networks based on themes/topics. First results and the extended applications of the method are discussed, which could lead to further insights into transmission processes of medieval miscellanies and, connected to these, of knowledge.

---

4 See for example Sarah Westphal-Wihl, *Textual Poetics of German Manuscripts, 1300–1500,* Studies in German Literature, Linguistics, and Culture (Columbia, SC, 1993); Michael Johnston and Michael Van Dussen, "Introduction: Manuscripts and Cultural History," in *The Medieval Manuscript Book. Cultural Approaches,* ed. Michael Johnston and Michael Van Dussen, Cambridge Studies in Medieval Literature 94 (Cambridge, 2015), 2–16; Nichols, "What Is a Manuscript Culture? Technologies of the Manuscript Matrix."

5 Following the terminology of Snijders, see note 1.

6 Differing titles for the same work at times makes identification difficult.

7 Tracing shared transmissions is potentially infinite: "Jede Mitüberlieferung einer Handschrift eröffnet eine eigene Textgeschichte, die wiederum häufig mit anderen Textgeschichten anderer Texte in dieser Handschrift verbunden sein kann." Freimut Löser, "Überlieferungsgeschichte(n) schreiben," in *Überlieferungsgeschichte transdisziplinär: Neue Perspektiven auf ein germanistisches Forschungsparadigma,* ed. Dorothea Klein, Horst Brunner, and Freimut Löser, Wissensliteratur im Mittelalter 52 (Wiesbaden, 2016), 15.

2. Networks of Texts

During my research on the German chronicle written by the Strasbourg cleric Jakob Twinger von Königshofen at the end of the fifteenth century,[8] I was confronted with a case of a particularly complex manuscript transmission.[9] Up to today, 128 manuscripts are known that contain the chronicle, wholly or in parts, and that were produced not only in Strasbourg, Twinger's home town, but as far away as Cologne, Augsburg, and Tyrol.[10] Around thirty of the manuscripts qualify as true, unedited copies, while in the large majority of the witnesses, the text differs in various ways:[11] abbreviated, augmented, updated, corrected, put in a different order, and more often than not combined with other texts, either with distinct boundaries marked by layout, headings, etc., or resulting in new compositions composed of several texts, where two or more were combined into new entities.[12]

The chronicle consists of six chapters, of which the last is an extensive index. While the first three chapters depict universal history,[13] chapters four and five narrate the past of the diocese and of the city of Strasbourg. The chronicle therefore covers many different interests and subjects: world history, the histories of secular and ecclesiastical rulers, as well as regional, diocesan and urban history. Furthermore, it contains a compressed knowledge tool in the form of the aforementioned index. Hence it is not very surprising that complete copies of the chronicle were often not necessary, or asked for, but that in some codices only the universal history was copied, for example – to serve as a basis for the insertion

8 Carl Hegel, ed., *Die Chroniken der oberrheinischen Städte. Straßburg 1,* Die Chroniken der deutschen Städte vom 14. bis ins 16. Jahrhundert 8 (Leipzig, 1870), 230–498; Carl Hegel, ed., *Die Chroniken der oberrheinischen Städte. Straßburg 2,* Die Chroniken der deutschen Städte vom 14. bis ins 16. Jahrhundert 9 (Leipzig, 1871).

9 See Ina Serif, *Geschichte aus der Stadt. Überlieferung und Aneignungsformen der deutschen Chronik Jakob Twingers von Königshofen,* Kulturtopographie des alemannischen Raums 11 (Berlin/Boston, 2020).

10 For an up-to-date list, see Ina Serif, "Der zerstreute Chronist. Zur Überlieferung der deutschsprachigen Chronik Jakob Twingers von Königshofen," *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte,* May 12, 2015, https://mittelalter.hypotheses.org/7063, last updated June 28, 2023. The entry in the medieval manuscript database *Handschriftencensus* currently records 115 entries, see https://handschriftencensus.de/werke/1906 [last accessed December 21, 2022].

11 Carl Hegel, who edited the chronicle in 1869/70, divided the transmission into three versions, A, B, and C, and based his edition on C, a version uniquely found in Twinger's autograph that burnt in 1870 in the Strasbourg library. For a discussion of Hegel's editorial decision regarding the prevalence of versions A and B in the existing manuscripts, see Serif, *Geschichte aus der Stadt,* 27–32.

12 Or 'scripta', following the terminology of Snijders, see note 1.

13 The first chapter spans from the Creation to Alexander the Great, the second and third give an account of the history of the Roman emperors, beginning with Caesar, and of the popes, starting with Peter.

of historiographical accounts of another town, substituting chapters four and five with local chronicles or annals.[14]

One way of tracing the transmission of this particular work is a text-based analysis. The co-occurrence of certain texts in several manuscripts hints towards intentional copying processes that reflect specific interests, not only of one individual scribe or commissioner. Detecting and tracing these occurrences can tell us more about reading interests and habits. Multiple occurrences of particular combinations can reveal connections that would go unseen if the content of a codex were not regarded as a whole, and it can tell us more about the migration of manuscripts. But apart from the pitfalls of working with manuscript descriptions mentioned above, a text-based analysis can face other difficulties. With respect to the transmission of the Twinger chronicle, the problem lies primarily in the sheer amount of testimonies. I built a database from existing manuscript descriptions, as well as from my own examinations, which contains the basic codicological information for all known textual witnesses – physical properties like writing surface, number of pages, and the dimensions of the codex, as well as its contents.[15] In the context of the transmission of the Twinger chronicle, nearly 500 different, distinct texts were identified within the 128 manuscripts, most of which appear only once in the corpus, with a few co-occurring in several codices. An attempt to analyse the patterns of textual transmission is rather unhelpful in this case, given the unique appearance of many texts (see Fig. 1).

In terms of concrete numbers, there are 437 single appearances of 489 texts, making up 89% of the total. This high percentage is to be expected and can be partially explained by the incompleteness of the data, due to the incoherence of the available manuscript descriptions mentioned above, and because not all codices could be analysed extensively to differentiate entries like "various prayers" or "several poems" into their constituent parts. If these could be split and potentially connected to other texts in the corpus, this would probably not change the general tendency, but rather point towards hitherto unknown connections in the manuscript transmission.

However, despite the impression of a lack of connections, some smaller clusters can be detected that share more than one text. The so-called *Konstanzer Jahrgeschichten* may serve as an example to illustrate potential insights, as

14 Good examples are the codices Freiburg im Breisgau, Universitätsbibliothek, Hs. 471 and Cologne, Historisches Archiv, Best. 7030 22. For a codicological overview, see https://handschriftencensus.de/13868 and https://handschriftencensus.de/12948, with further literature.

15 The data can be downloaded from Zenodo: https://doi.org/10.5281/zenodo.7469112 or via the GitHub repository: https://github.com/wissen-ist-acht/twinger_chronicle_mss.

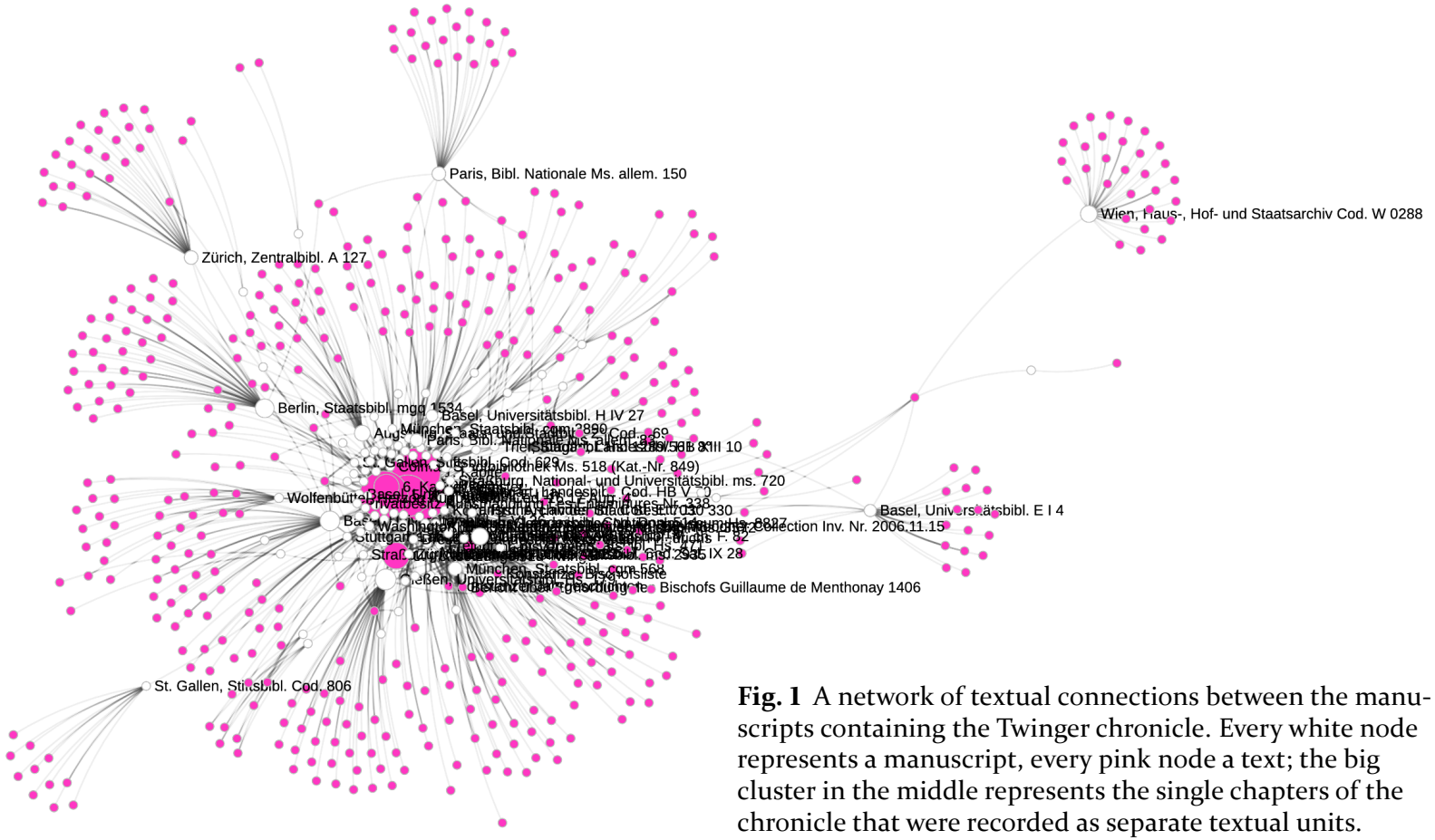

Fig. 1 A network of textual connections between the manuscripts containing the Twinger chronicle. Every white node represents a manuscript, every pink node a text; the big cluster in the middle represents the single chapters of the chronicle that were recorded as separate textual units.

eISSN: 2535-8863
DOI: 10.25517/jhnr.v9i1.136

well as the pitfalls of constructing a purely text-based network. The *Jahrgeschichten* are a short annalistic record in German consisting of notes, mainly about events in the city of Constance. The notes begin with the year 1256, when *bruoder Berchtold* preached in Constance for the very first time, and end, in most of the manuscripts, with 1388, reporting a huge fire in Constance and the neighbouring town of Stadelhofen. Twelve manuscripts that contain the *Jahrgeschichten* are known, eleven of which also contain the Twinger chronicle completely or in part.[16] A closer look at the manuscripts shows that they were not only produced in the city of Constance, where an interest for the *Jahrgeschichten* is rather self-evident, but also in Augsburg, some 200 kilometres northeast of it (see Fig. 2).[17]

Earlier research has tended to over-interpret the co-occurrence of this account with another historiographical text like the Twinger chronicle, leading to generalizations such as the following:

> Die in der Bischofsstadt Straßburg geschriebene Chonik hat in Konstanz ein breites Interesse gefunden. Es sind zehn [elf, I. S.] Handschriften bekannt, in denen an Twingers Text eigene Lokalnachrichten zur Konstanzer Geschichte angeschlossen wurden.[18]

16 Basel, UB, Cod. E VI 26; Dürnstein, Regularkanonikerstift, s.n. [now lost]; Freiburg im Breisgau, UB, Hs. 471; Gotha, FB, Cod. Chart. A 158; Heidelberg, UB, Cod. Sal. IX 28; Heidelberg, UB, Cpg 475; Karlsruhe, BLB, Cod. Don. 513; Munich, BSB, Cgm 567; Munich, BSB, Cgm 568; St. Gallen, Stiftsbibl., Cpg 630; Strasbourg, BNUS, ms. 5457. The codex Constance, Stadtarchiv, A I 1 only contains the *Jahrgeschichten,* without the passages from the Twinger chronicle. On the *Jahrgeschichten,* see Klaus Graf, "Die verschollene Twinger-Handschrift aus dem Regularkanonikerstift Dürnstein," *Archivalia,* March 27, 2013, https://archivalia.hypotheses.org/6941; Ina Serif, "Konstanzer Jahrgeschichten," in *Encyclopedia of the Medieval Chronicle, 2nd Online Edition,* ed. Graeme Dunphy and Christian Bratu (Leiden/Boston, 2016), http://dx.doi.org/10.1163/2213-2139_emc_SIM_001450. For the text of the single entries, see Franz Josef Mone, ed., *Quellensammlung der badischen Landesgeschichte 1* (Karlsruhe, 1848), 302–303; Franz Josef Mone, ed., *Quellensammlung der badischen Landesgeschichte 3* (Karlsruhe, 1863), 509; Gustav Scherrer, *Kleine Toggenburger Chroniken. Mit Beilagen und Erörterungen* (St. Gallen, 1874), 93–97.

17 The origin of two of the codices in Strasbourg and Basel is explained by manuscript migration: the *Jahrgeschichten* were added at a later time, after the codices had left their place of origin – another fact which complicates transmission analyses. For three manuscripts, we do not have enough evidence (yet) for a precise localization.

18 "The chronicle that was produced in the episcopal city of Strasbourg arouse interest in Constance. Ten [eleven, I. S.] manuscripts are known in which own local news concerning the history of Constance were inserted after Twinger's work." See Eugen Hillenbrand, "Gallus Öhem, Geschichtsschreiber der Abtei Reichenau und des Bistums Konstanz," in *Geschichtsschreibung und Geschichtsbewußtsein im späten Mittelalter,* ed. Hans Patze, Vorträge und Forschungen 31 (Sigmaringen, 1987), 734.

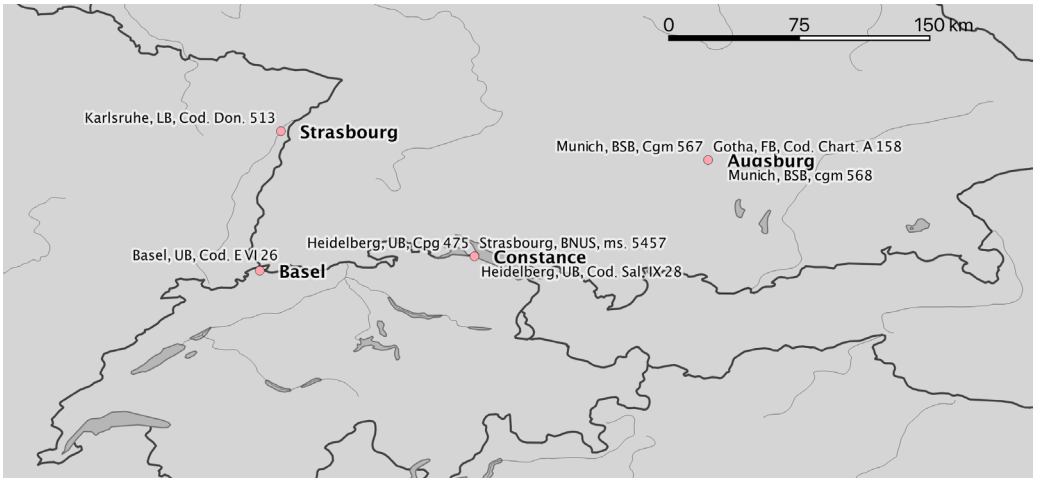

Fig. 2 Known places of production for eight of the manuscripts that contain the Twinger chronicle and the *Konstanzer Jahrgeschichten*.

When we look at the eleven codices in question, we see that they indeed share several texts – apart from the chapters of the Twinger chronicle (labelled here as *1. Kapitel, 2. Kapitel* and so forth) and the *Jahrgeschichten,* they include an account of the murder of the bishop of Lausanne, Guillaume of Menthonay, and a list of the bishops of Constance (Fig. 3).

Apart from the shared transmission, it is just as interesting to investigate which texts are *not* shared between the manuscripts, and to explore whether this allows for further assumptions or insights with respect to the compilation processes and manuscript migration. To be able to compare the different texts, I attributed tentative genres to them.[19] This reveals remarkable differences between some of the copies that are in need of explanation (Fig. 4).

While the codex Munich, BSB, Cgm 567 was probably in Hillenbrand's mind when he stated that the Twinger chronicle provokes interest in Constance and that local news concerning the history of the town were inserted after Twinger's

19 Genre attribution is always interpretive, and different levels of description are applied, such as structure or content. Several ontologies and references exist, all with their own advantages and disadvantages, e.g., the database *Geschichtsquellen des Mittelalters* that classifies every listed work, but without elaborating on the scheme, see https://www.geschichtsquellen.de/filter?filter=gattung. For my sample, I used 28 genres, without following a specific ontology: account, annals, chronicle, confession treatise, contract, didactic poem, directory, episcopal history, exemplum, family history, legend, letter, list, medical treatise, notes, novel, parody, poetry, proverb, reformatory account, regional history, royal legislation, treatise, universal history, urban history, vocabulary, and war history.

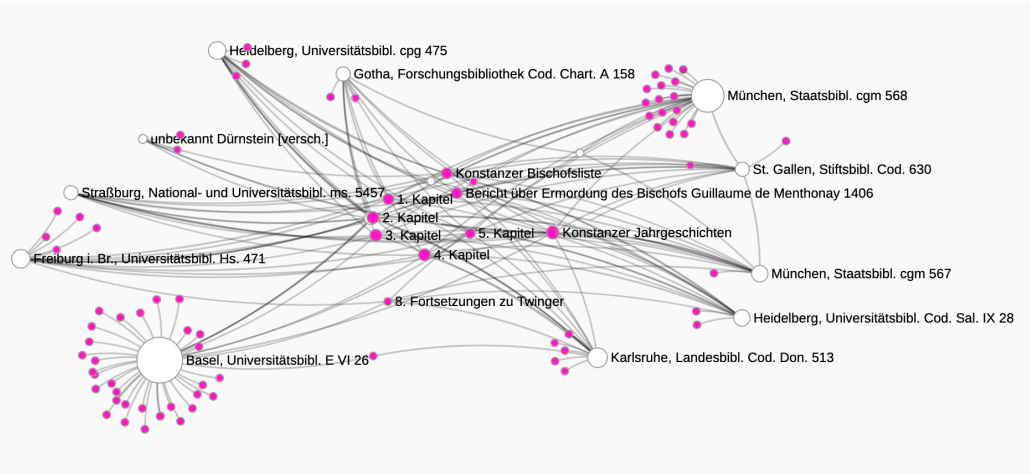

Fig. 3 Shared textual tradition in eleven manuscripts that contain, besides the Twinger chronicle, the *Konstanzer Jahrgeschichten,* a list of the bishops of Constance and an account of the murder of Guillaume of Menthonay, bishop of Lausanne, in 1406.

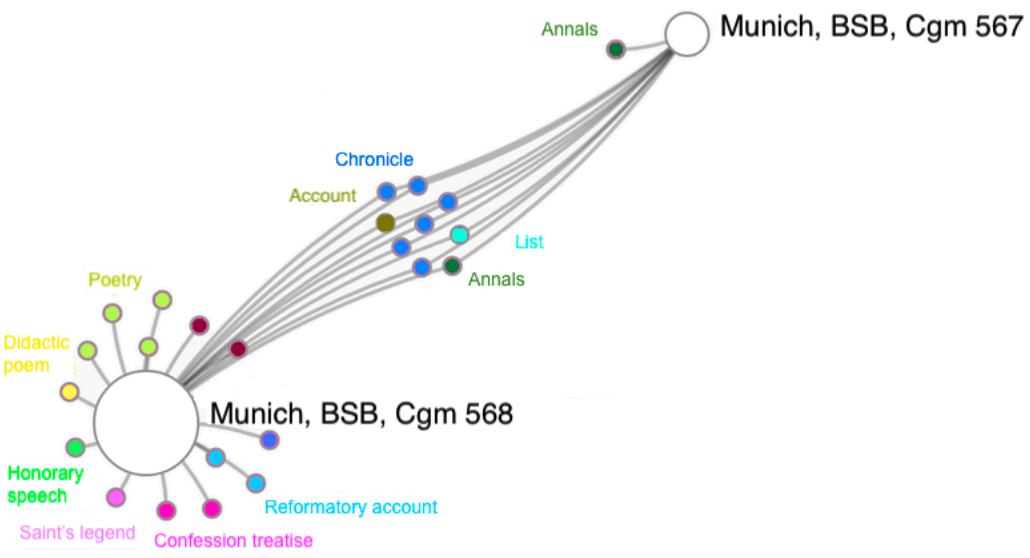

Fig. 4 Content of the codices Munich, BSB, Cgm 567 and Munich, BSB, Cgm 568, including genres.

work, the other compilation, Munich, BSB, Cgm 568, provides a different picture: apart from historiographical texts, we find poems, didactic poems, reformational accounts, confession treatises, and a legend of a saint.[20] A look at the actual texts reveals an interesting compilation in which universal and regional history was enhanced with religious and literary texts: after universal, regional and Strasbourg history, mediated by the Twinger chronicle, there follow two accounts about events of another city with the *Konstanzer Jahrgeschichten* and a list of the bishops of Constance. Afterwards, a list of the bishops of Augsburg is inserted, followed by the legend of St. Ulrich, one of the first bishops of Augsburg, leading to pieces with a religious focus, supplemented by prayers and treatises in Latin and German. The *Reformatio Sigismundi* and the *Reformatio Friderici* point towards the realm of literature, as do the poems of Thomas Prischuch and those of Jörg Zobel, added later. Thus, instead of a focus on the history of Constance, we find the opposite: the single texts mostly cover events in or persons from Augsburg – such as the list of bishops or the legend of the saint – or derive from Augsburg-based authors, like the poems of Thomas Prischuch. During the process of compilation, the original structure of the template,[21] namely the Twinger chronicle, the *Jahrgeschichten* and the list of bishops, was copied, but the addition of texts from quite different genres reduced the historiographical character of the collection. The broadening of the subjects also widened the potential readership, possibly resulting in an increased appeal of the production of compilations that were not purely historiographical in content. The question thus arises of whether a genre-based approach, and an attempt to construct networks from genres instead of texts, can help to better understand medieval manuscript production and make up for the insufficiencies of a text-based analysis.

3. Networks of Genres

Comparing manuscripts by looking at the genres of the texts they contain instead of the texts themselves can offer new insights, showing not only connections between manuscripts that would otherwise remain unseen, but also pointing towards interesting compilations from a programmatic point of view. There are two possible approaches to such a comparison: by attributing a genre to every text in a

20 For an overview of the codices, see https://handschriftencensus.de/9896 (Munich, BSB, Cgm 567) and https://handschriftencensus.de/6173 (Munich, BSB, Cgm 568).

21 For the identification of Munich, BSB, Cgm 567 as a template for Munich, BSB, Cgm 568, see Hegel, *Die Chroniken der oberrheinischen Städte. Straßburg 1,* 220. The main scribe is Johannes Erlinger, who may have produced the codex for himself; see Karin Schneider, "Berufs- und Amateurschreiber. Zum Laien-Schreibbetrieb im spätmittelalterlichen Augsburg," in *Literarisches Leben in Augsburg während des 15. Jahrhunderts,* ed. Johannes Janota and Werner Williams-Krapp, Studia Augustana 7 (Tübingen, 1995), 20–21. Apart from Erlinger's hand, there is a short addition by Konrad Bollstatter, a well-known Augsburg scribe, and the poems of Zobel were added at a later stage by an unknown hand.

manuscript, as has been shown for the Munich codices (see Fig. 4); or by classifying a manuscript as a unit, based on its textual content as a whole.

An attempt to apply the first approach on the entire sub-corpus of the eleven manuscripts that contain both the Twinger chronicle and the *Konstanzer Jahrgeschichten* shows some dominant genres (see Fig. 5).

If we compare this genre-based network with the text-based network above (Fig. 3), some compilations appear more coherent with regards to an underlying concept.[22] The codices Heidelberg, UB, cpg 475 (upper left), Freiburg im Breisgau, UB, Hs. 471 (on the left), and Karlsruhe, BLB, Cod. Don. 513 (lower right), for instance, seem to contain mainly historiographical texts, whereas Basel, UB, E VI 26 (lower left) and Munich, BSB, Cgm 568 (upper right) showcase a wider variety of genres. This kind of analysis seems to be possible for a small corpus, or miscellanies with few texts; it is not endlessly scalable, at least not so long as genre attribution is based on manual classification – not to mention the subjective nature of such a classification.

For an individual researcher, the second approach, classifying miscellanies as a whole, might be more feasible. The underlying hypothesis is that compilations were put together following some kind of concept, often combining works of the same genre. An exploratory analysis carried out by Gustavo Riva on the basis of 26,000 manuscripts containing Middle High German texts supports this assumption.[23] Working with the data of the *Handschriftencensus,* an inventory of the manuscript tradition of medieval German language texts, he constructed a network of shared manuscript transmission that shows clusters of texts that can be assigned to single genres. These clusters are of course fuzzy at the borders, but they show some broader tendencies, like the frequent combination of texts with similar genres in multiple-text manuscripts. This kind of analysis also shows that some texts are likely to fit into any kind of context, independent of the genre(s) of

22 We still lack a consistent terminology for manuscripts containing more than one text; while 'miscellany' is probably the least specific, terms like 'one-volume libraries' or 'multiple-text manuscripts' are in use, without clear definitions, and without referring to chronological aspects of the production, nor to structural or material characteristics. For recent reflections and studies see Michael Friedrich and Cosima Schwarke, eds., *One-Volume Libraries. Composite and Multiple-Text Manuscripts,* Studies in Manuscript Cultures 9 (Berlin; Boston, 2016); Marilena Maniaci, "Miscellaneous Reflections on the Complexity of Medieval Manuscripts," in *Collecting, Organizing and Transmitting Knowledge. Miscellanies in Late Medieval Europe,* ed. Sabrina Corbellini, Giovanna Murano, and Giacomo Signore, Bibliologia: Elementa ad Librorum Studia Pertinentia 49 (Turnhout, 2018), 11–22; Alessandro Bausi, Michael Friedrich, and Marilena Maniaci, eds., *The Emergence of Multiple-Text Manuscripts,* Studies in Manuscript Cultures 17 (Berlin/Boston, 2019).

23 Gustavo Fernández Riva, "Network Analysis of Medieval Manuscript Transmission. Basic Principles and Methods," *Journal of Historical Network Research* 3 (2019): 30–49.

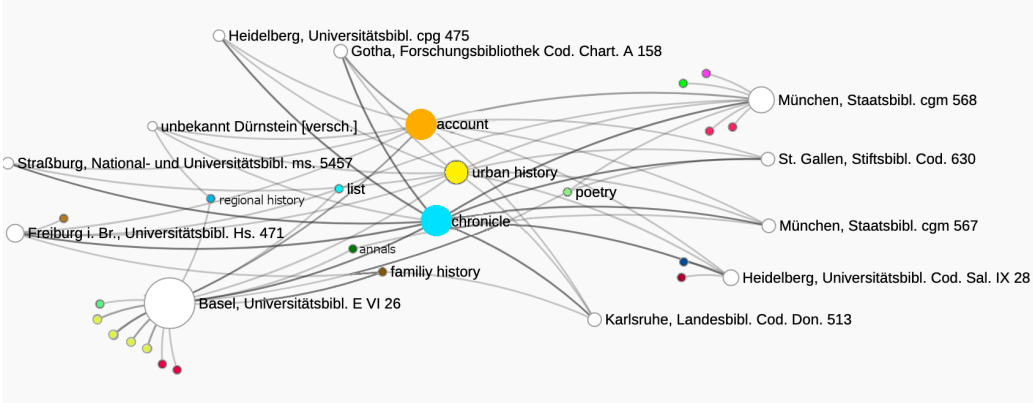

Fig. 5 Shared genre tradition in eleven manuscripts that contain, besides the Twinger chronicle (chronicle), the *Konstanzer Jahrgeschichten* (urban history), a list of the bishops of Constance (list), and an account of the murder of Guillaume of Menthonay, bishop of Lausanne, in 1406 (account).

the other texts.[24] But again, we encounter the question of classification: the attribution of a genre to a text is all but undisputed, making it simultaneously a productive yet hindering category.[25]

And while the classification of an (abstract) work or its textual manifestation is difficult enough, another difficulty arises with regards to the classification of miscellanies as a unit, a problem that has not yet been satisfyingly addressed.[26] More often than not, one main genre is chosen as a designator for the whole context, resulting in a "clash between genre study and codicology."[27] Miscellanies are more than the sum of their parts, but classifying them is very difficult – if not impossible[28] – and this is all the more true for compiled manuscripts that do not immediately appear homogeneous with respect to the (ascribed) genres of the

24 Ibid., 46.
25 Westphal-Wihl, *Textual Poetics of German Manuscripts, 1300–1500,* 8.
26 "But the vast majority of manuscripts have a miscellaneous character that defies the concept of genre as a principle of identity and separation." Ibid., 9.
27 Ibid. For an attempt to classify manuscripts by type to conduct a network analysis, see Octave Julien, "Délier, lire et relier. L'Utilisation de l'analyse réseau pour construire une typologie de recueils manuscrits de la fin du Moyen Âge," *Hypothèses* 19, no. 1 (2016): 211–24, doi:10.3917/hyp.151.0211. Julien groups the (mostly multi-text) manuscripts into ten categories: religion, moral literature, profane literature, history and politics, jurisprudence, practical texts, scientific texts, domestic and technical literature, and encyclopedias. He also refers to sub-categories, without further explanation.
28 "Sammelhandschriften sind mehr als die Summe der Einzelteile, aber sie in diesem integralen Sinn in den Blick zu nehmen, ist beinahe unmöglich." Jürgen Wolf, "Sammelhandschriften – mehr als die Summe ihrer Einzelteile," in *Überlieferungsgeschichte transdisziplinär: Neue Perspektiven auf ein germanistisches Forschungsparadigma,* ed.

containing texts or an (assumed) programmatic arrangement – as has been illustrated here for the codex Munich, BSB, Cgm 568.[29]

Thus, the questions arise: how should we deal with the codicological context of a work, and how can we make this context productive for analyses of production processes of reader's interests, of knowledge spreading, without perpetuating narrowing classifications, or replacing them with new constrictions? Is there something like a "unifying purpose"[30] of a compiled manuscript, and how could it be detected?[31]

4. Networks of Topics

As the manual classification of the content of compiled manuscripts poses several methodological problems, a computational approach may serve as an alternative. This can help to compensate for incomplete or inaccurate manuscript descriptions, in order to overcome the subjectiveness of classifications and allow scalability and therefore applicability for the different source corpora. The attempt to connect manuscripts on the basis of computationally calculated topics and its suitability for medieval miscellanies will be discussed in the following, again using codices from the Twinger chronicle as an example.

Dorothea Klein, Horst Brunner, and Freimut Löser, Wissensliteratur im Mittelalter 52 (Wiesbaden, 2016), 80.

29 "Das gilt besonders für die Textformationen, die sich nicht auf einen Nenner der Art 'enzyklopädische Sammlung', 'historiographische Kompilation', 'Liedersammlung', 'Bibelkompendium' oder 'Legendar' bringen lassen, und die als ein mehr oder weniger zufälliges Sammelsurium von Texten ohne bedeutungstragende Ordnungs- und Organisationsstruktur erscheinen. Es fehlen außerdem systematische Überlegungen zur Historisierung und zur Spezifik des Mediums 'Sammelhandschrift'. Statt auf generelle Funktionsweisen von Textzusammenstellungen wird der Fokus bislang auf einzelne Bücher und deren Rezipienten gerichtet, auf die Programmatik individueller Textsammlungen oder auf das historische Verständnis von Gattungen und Genres, das man aus den Textkombinationen meint ableiten zu können." See Diana Müller, *Textgemeinschaften. Der "Gregorius" Hartmanns von Aue in mittelalterlichen Sammelhandschriften* (Frankfurt a.M., 2013), 42, http://publikationen.ub.uni-frankfurt.de/frontdoor/index/index/docId/30069.

30 Stephen G. Nichols and Siegfried Wenzel, eds., "Introduction," in *The Whole Book: Cultural Perspectives on the Medieval Miscellany, Recentiores.* Later Latin Texts and Contexts (Michigan, 1996), 6.

31 The temporal aspect and the evolving character of manuscripts with regards to their content and structure, and therefore evolving "unifying purposes", are left out of this analysis; we still have to keep in mind that the "manuscript was in constant flux, always with the potential to be reshaped by its current owner." See Johnston and Van Dussen, "Introduction: Manuscripts and Cultural History," 5.

I chose topic modeling as a means to quantitatively approach texts.[32] Rather than merely counting word frequencies, like tf-idf,[33] the underlying idea behind topic modeling is that words that appear in the same context have the same, or a similar, meaning. Therefore, not only the frequency, but also the distribution of words in a document, or a corpus, is counted, using statistical methods. Depending on the distributions, topics are inferred, each of them consisting of a list of words that appear together in a statistically significant way. For my case study, I used the Dariah Topics Explorer,[34] software that is based on the statistical model Latent Dirichlet Allocation (LDA).[35] As this is a GUI tool, not all parameters that the model is based on can be changed, but it is a good starting point to determine rather quickly whether an analysis with topic modeling is a useful approach for a specific corpus. For any further analysis, I would recommend using programs that allow for complete control of all steps.[36]

---

32 For a useful overview of introductory texts, more technical articles and research projects that use topic modeling, see Scott B. Weingart, "Topic Modeling for Humanists: A Guided Tour," July 25, 2012, http://scottbot.net/topic-modeling-for-humanists-a-guided-tour/. Anne Purschwitz applied topic modeling to historical journals of the Enlightenment, attempting to discover (networks of) scholarly discourses, see Anne Purschwitz, "Netzwerke des Wissens – thematische und personelle Relationen innerhalb der Halleschen Zeitungen und Zeitschriften der Aufklärungsepoche (1688–1818)," *Journal of Historical Network Research* 2 (December 3, 2018): 109–42. For some critical remarks on the data basis of the construction of networks based on the result of topic modeling see Scott B. Weingart, "Topic Nets," November 10, 2012, http://scottbot.net/topic-nets/.

33 Term frequency–inverse document frequency. For a broader discussion of this statistical measure, see Stephen Robertson, "Understanding Inverse Document Frequency: On Theoretical Arguments for IDF," *Journal of Documentation* 60, no. 5 (January 1, 2004): 503–20, doi:10.1108/00220410410560582.

34 Available at https://github.com/DARIAH-DE/TopicsExplorer; Steffen Pielström, Severin Simmler, and Thorsten Vitt, "Topic Modeling with Interactive Visualizations in a GUI Tool," *Proceedings of the Digital Humanities Conference Utrecht 2019,* n.d., https://dev.clariah.nl/files/dh2019/boa/0637.html. For a short tutorial in German see Mareike Schuhmacher, "DARIAH Topics Explorer," *ForTEXT. Literatur digital erforschen,* accessed October 29, 2021, https://fortext.net/tools/tools/dariah-topics-explorer.

35 The model was introduced by David Blei, Andrew Ng and Michael I. Jordan for use in textual studies, see David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet Allocation," *The Journal of Machine Learning Research* 3 (March 1, 2003): 993–1022. David M. Blei, "Probabilistic Topic Models," *Communications of the ACM* 55, no. 4 (April 2012): 77–84. For a concise (and humanist approved) explanation of LDA and Gibbs sampling see Ted Underwood, "Topic modeling made just simple enough," July 4, 2012, https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/.

36 Gensim and Mallet are two programs that enable adjustment: Gensim is a Python library, while Mallet is based on Java (Dariah's Topics Explorer uses Mallet). A comparison of the results of the two programs, which use different sampling methods for different corpora, one of them the subset of Twinger manuscripts used here, can be found in Tobias Hodel, Dennis Möbus, and Ina Serif, "Von Inferenzen und Differenzen. Ein Vergleich von Topic-Modeling-Engines auf Grundlage historischer Korpora," in *Von Menschen und Maschinen. Mensch-Maschine-Interaktionen in digitalen Kulturen,* ed. Selin Gerlek et al. (Hagen 2022), 181–205. We are currently working on a follow-up paper that reflects in

The calculation of topics is based on the full text of a document, independent of text boundaries, how it has changed hands, or the different stages of production of each individual text. This allows for an analysis of the content of manuscripts without the need for prior manual inspection; this means that the method is potentially usable on very large corpora, given that their full text is provided in a machine-readable format.[37]

The corpus used for this proof of concept consists of seven multiple-text manuscripts, all containing the Twinger chronicle, three of which are part of the *Jahrgeschichten* corpus:[38] Dresden, UB, Mscr. F 98; Freiburg im Breisgau, UB, Hs. 471 (with *Jahrgeschichten*); Heidelberg, UB, Cpg 116; Heidelberg, UB, Cpg 475 (with *Jahrgeschichten*); Munich, BSB, Cgm 568 (with *Jahrgeschichten*); Stuttgart, LB, HB V 22; Wolfenbüttel, HAB, Cod. 16.17. I performed handwritten text recognition (HTR) on all seven manuscripts, using the software Transkribus and some of its generic, publicly available models.[39] I did not perform an elaborate post

---

detail on the methods and concepts used for topic modeling, such as the different steps of preprocessing – lower casing, removal of punctuation, chunking, etc. – and the adjustment of several parameters. Chunking in particular – the cutting of documents into equal parts – seems to compensate for the varying length of the medieval manuscripts.

37 Christof Schöch applied topic modeling to French Classic and Enlightenment literature and to texts of Arthur Conan Doyle respectively, attempting to model genre see Christof Schöch, "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama," *Digital Humanities Quarterly* 11, no. 2 (2017); Christof Schöch, "Computational Genre Analysis," in *Digital Humanities for Literary Studies: Methods, Tools, and Practices,* ed. James O'Sullivan (College Station, TX, 2020), 219–31.

38 In a best-case scenario, the test corpus built for this trial would have consisted of the eleven manuscripts that share the Twinger chronicle and the *Konstanzer Jahrgeschichten,* because the comparison of the three different approaches discussed here would have been more exact. But some constraints led to a slightly different composition, mainly the lack of digital copies (Dürnstein, Regularkanonikerstift, s.n. [now lost]; Gotha, FB, Cod. Chart. A 158 [digitized very recently]; Munich, BSB, Cgm 567; St. Gallen, Stiftsbibl., Cpg 630; Strasbourg, BNUS, ms. 5457). Another reason was the rather unpromising results of handwritten text recognition (HTR) using existing models for certain manuscripts.

39 To perform the text recognition on the manuscripts, I received free credits from Transkribus, applying for their scholarship program (https://readcoop.eu/transkribus/scholarship) – thanks again! The models used were "German Kurrent XVI–XVIII M1", "Thun Missiven M3", "Medieval Scripts M2" and "Charter Scripts XIII–XV M4". (All of these are based on the CitLab HTR/HTR+ engine, which is no longer supported by Transkribus; while there are models based on the engine PyLaia that fit the different writings in the used manuscripts, the transcription output might differ from the output achieved with the other engine.) Playing around with the different models is worth the while, in particular with those covering longer time periods. The accuracy might suffer a bit, but especially for codices that consist of several texts by various hands, even some that are decades or centuries apart, in many cases there is no need to apply different models on different parts of the manuscript, so the time saved makes up for the lower accuracy.

processing on the resulting text, for two reasons.[40] Firstly, I would like to provide a proof of concept of this approach for larger corpora, and the necessity of lengthy HTR correction or other forms of data cleaning would preclude scalability.[41] Secondly, lemmatization or normalization runs the risk of altering the resulting topics in an unwanted manner: the peculiarities of premodern writing, without fixed spelling and dialectal variations, are already part of linguistic analyses to discover relationships between manuscripts; how this shows in a topic model will be discussed in a moment.[42]

To be able to evaluate the results on a slightly larger scale, the corpus was enlarged with some printed editions of several medieval texts, to see if the resulting topics differ from those with uncorrected texts. I also added the edition of the Twinger chronicle, to see how much the topics of the single manuscripts would be related to those of the edited text. Thus, in addition to the codices mentioned above, the printed text of the *Chronik der Eidgenossenschaft* by Petermann Etterlin, the *Oberrheinische Chronik,* the *Konzilschronik* by Ulrich Richental, the *Leben des heiligen Ulrich* by Albert von Augsburg, and the Twinger chronicle were also used for topic analysis, the latter being divided into two parts, the first containing chapters one, two and three, and the second with chapters four, five and six.[43] In none of the editions did any normalization take place, so the difference

40 Apart from stripping strings which contained information about the digitizing institution: many digitized images are marked with a copyright sentence, which mentions the holding library; this text was also recognized during recognition, but could be easily detected and deleted. Also, diacritics were dissolved: during text recognition, words with diacritics were often split into two strings, e.g., "brů" and "der"; dissolution resulted in one string, e.g., "bruoder".

41 For a meta study on the impact of OCR errors see Stephen Mutuvi et al., "Evaluating the Impact of OCR Errors on Topic Modeling," in *Maturity and Innovation in Digital Libraries,* ed. Milena Dobreva, Annika Hinze, and Maja Žumer, Lecture Notes in Computer Science (Cham, 2018), 3–14, doi:10.1007/978-3-030-04257-8_1. While there is a measurable impact of OCR errors on the output of topic model analyses, this impact is relatively small overall. Apparently, it doesn't affect the average coherence score between the models too much, Ibid., 12.

42 The data used for the analysis – the txt-files as well as the output of the topic modeling – are available at: https://github.com/wissen-ist-acht/tm_data.

43 Eugen Gruber, ed., *Petermann Etterlin. Kronica von der loblichen Eydtgnoschaft, Jr harkommen und sust seltzam stritten und geschichten,* Quellenwerk zur Entstehung der Schweizerischen Eidgenossenschaft 3, 3 (Aarau, 1965); Franz Karl Grieshaber, *Oberrheinische Chronik: Älteste bis jetzt bekannte, in deutscher Prosa* (Rastatt, 1850); Thomas Martin Buck, ed., *Chronik des Konstanzer Konzils 1414–1418 von Ulrich Richental: historisch-kritische Edition. Band 1: A-Version,* vol. XLIX, 1–3, Konstanzer Geschichts- und Rechtsquellen (Ostfildern, 2020); Karl-Ernst Geith, ed., *Albert von Augsburg: Das Leben des heiligen Ulrich,* Quellen und Forschungen zur Sprach- und Kulturgeschichte der germanischen Völker, n.F. 39 (163) (Berlin/New York, 1971); Hegel, *Die Chroniken der oberrheinischen Städte. Straßburg 1;* Hegel, *Die Chroniken der oberrheinischen Städte. Straßburg 2.* While the text of the *Konzilschronik* (Aulendorfer version), the *Kronica* and *Das Leben des heiligen Ulrich* were digitally available, I performed text

between the manuscripts lies mainly in the greater accuracy of the transcription. Some texts were already available as a text file, while for others I had to perform OCR on the digitized books. Regarding post processing, the same applies as for the manuscripts: I did not edit the text recognized, I merely removed the superfluous information derived from the digitization process.[44]

Usually, text analysis methods like topic modeling, where occurrences of tokens are counted, use a list of stop words, i.e., words that should be excluded from the analysis because they appear often, but do not carry much (semantic) meaning, yet possess syntactical or grammatical functions. Prebuilt lists exist, also for premodern languages; however, these still need to be adapted and enlarged upon, mainly because of different ways of spelling, like "und", "unde", "unnd", "vnnd", "unnt", etc. for "and". For my analysis, I used the stop word list for Middle High German, provided by the Classical Language Toolkit,[45] which I extended during analysis. Topic modeling is, like many other approaches to analysing texts, iterative, meaning that the findings of a first inspection can be used to improve the results – for example, the topics that were detected in a first round led to the exclusion of several words for the next round by adding them to the stop list (see Fig. 6).

After several rounds of modeling and exclusion of more stop words, increasing the amounts of topics from ten to twenty-five and the number of iterations of the model to 10,000, the results looked more nuanced (see Fig. 7).

Aside from a list of topics, the Topic Explorer also offers a document-topic-matrix that represents a network of topics. This shows the prevalence of a topic in a document using saturation: the lighter a field is, the less important is the topic, or the less common are the words of this specific topic within a document (see Fig. 8).

A closer look at this matrix provides us with three different kinds of results (Fig. 9): first, the fact that the fifth topic, "strosburg, stat, bischof" is more prevalent in the second part of the Twinger chronicle (namely chapters four and five) is not surprising from what we already know about the content, nor is the dominance of the twelfth topic "künig, bobest, rome" in the first part (chapters one to

---

recognition for the remaining texts. The results are slightly better for modern print, i.e., not Gothic type, but still very decent for the latter.

44 I also did not delete the critical apparatus. In all the editions used here, the quantity of the edited text was easily sufficient to outweigh these remainders.

45 This Python library performs natural language processing, especially for premodern languages. At the moment, it is available for nineteen languages. See cltk.org for the package and its documentation.

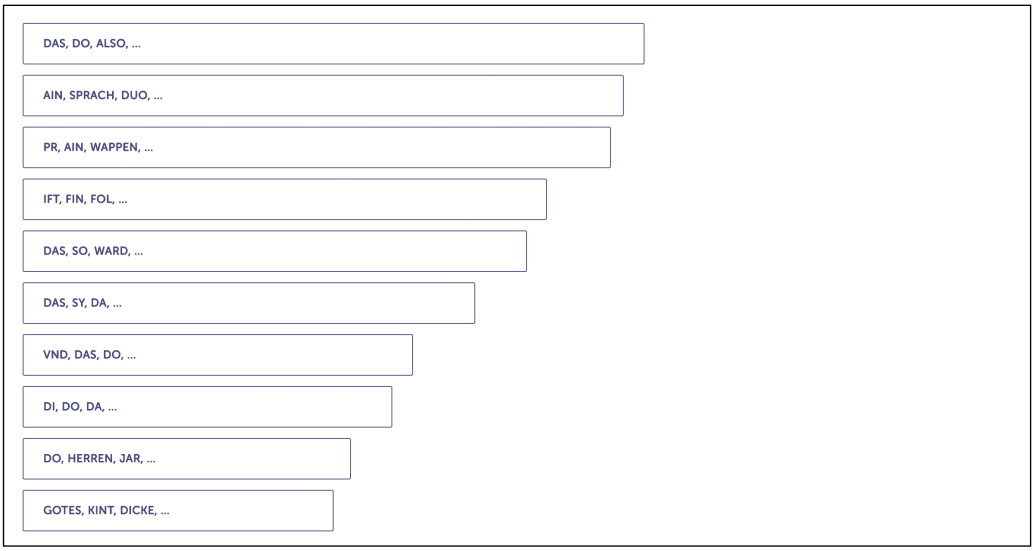

Fig. 6 List of ten topics, showing the three most common words for each topic (out of fifteen in total that are listed in the results). One side effect of omitting post processing is already visible, the incorrect recognition of "ſ" as "f" instead of "s".

Fig. 7 List of 25 topics, showing the three most common words for each topic. The skipped normalization and lemmatization are clearly visible, showing "kùng", "kùnig", "künig", "kung", "könig", "küning" as variations of today's "König" (king).

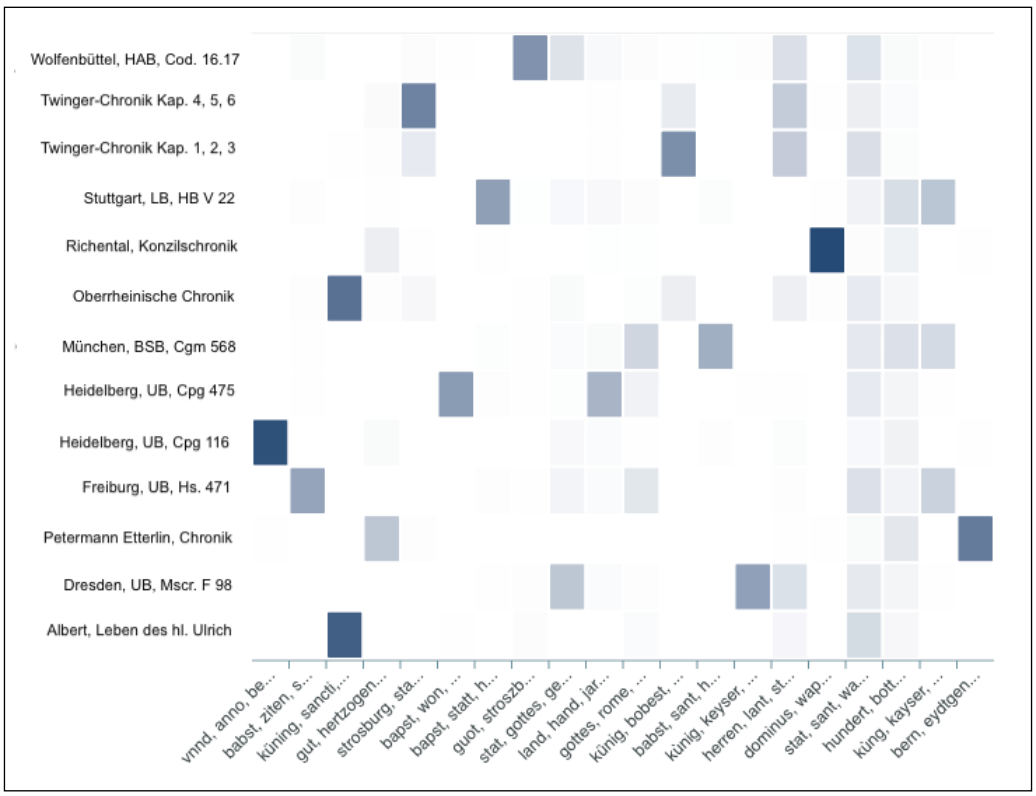

Fig. 8 Document-topic-matrix/topic network with weighed prevalence of each topic (13 documents, 20 topics).

three, green frames).[46] Here we can also see that topic modeling does not overly depend on lemmatization or "correct" spelling according to modern dictionaries. Second, we can confirm existing knowledge from a close reading of the manuscripts: the two codices Freiburg im Breisgau, UB, Hs. 471 and Munich, BSB, Cgm 568 share topic 11 with Heidelberg, UB, Cpg 475, and topic 19 with Stuttgart, LB, HB V 22 (orange frames). For the first group, we already know that the three miscellanies share several texts that deal with Constance, among them the *Konstanzer Jahrgeschichten.* In the second group, the Stuttgart codex does not contain the *Jahrgeschichten,* but two other texts that are concerned with the history of Constance: the *Konzilschronik* by Ulrich Richental and the *Konstanzer Chronik* by Gebhard Dacher. Here we get closer to the initial idea of identifying rela-

46 Unfortunately, there is no comprehensive visualization of the matrix with a list of words for the single topics. The overall trend should be visible, even through two or three words that are shown in the figure, and which I refer to in the text. A complete list of the words for each topic can be found in the repository: https://github.com/wissen-ist-acht/tm_data/.

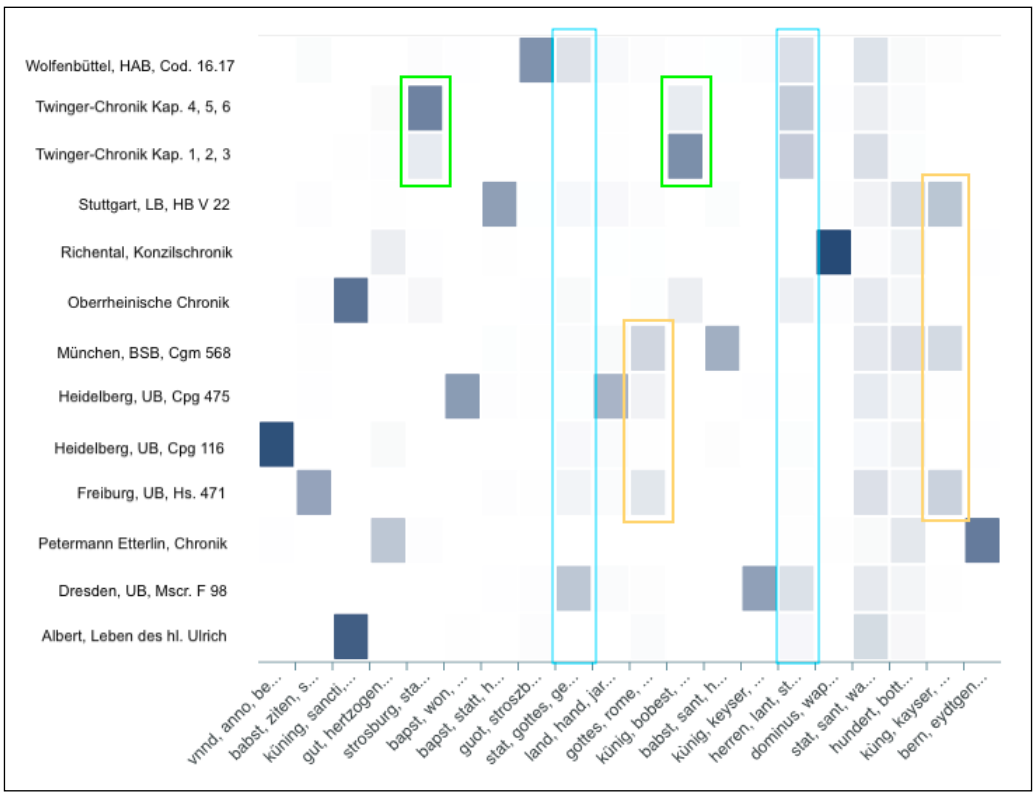

Fig. 9 Document-topic-matrix with weighed prevalence of a topic (13 documents, 20 topics), and special attention to topics five, nine, eleven, twelve, fifteen and nineteen.

tions between manuscripts through topics, without having to read them or rely on available descriptions. A third kind of finding is somewhat surprising: the two manuscripts Wolfenbüttel, HAB, Cod. 16.17 and Dresden, UB, Mscr. F 98 that share topics 9 and 15 do not have anything in common on a textual level – apart from the Twinger chronicle, of course (blue frames). The Dresden codex contains several texts that are concerned with the Burgundian War, whereas the Wolfenbüttel manuscript collects lyric, prayers, and cooking recipes.

The interpretability of such a matrix correlates with the number of documents. If we wanted to obtain a first impression of the relations between manuscripts in a much larger corpus, the visualization of the results as a network is helpful. However, for this small sample, we also get a nice impression of the connected codices (see Fig. 10).[47]

47 A CSV file of the document similarities as part of the Topics Explorer export was the basis for the network creation – many thanks go to Gustavo Riva for showing me how to

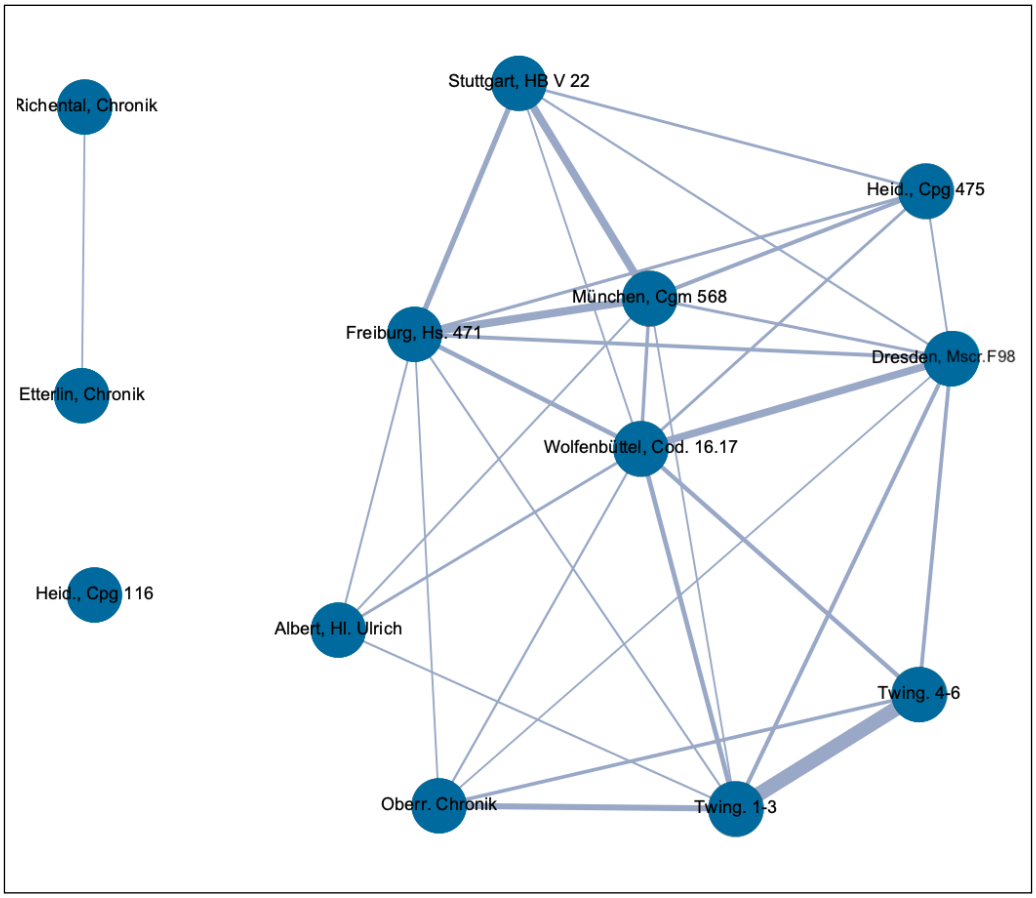

Fig. 10 Network of documents with weighed prevalence of a topic. Visualization created with Gephi.

As in the document-topic-matrix, we see stronger connections between the two parts of the Twinger chronicle (lower right), between the manuscripts with a focus on Constance (Freiburg im Breisgau, Stuttgart, Munich and Heidelberg, Cpg 475, at the top), as well as between the codices from Wolfenbüttel and Dresden (in the center). One result that is more visible in this kind of visualization is

do this! In this visualization, the edge weight ranges between 0.5 and 0.8. Higher and lower weights have been filtered out. Without this filter, a strong connection between the *Leben des heiligen Ulrich* and the *Oberrheinische Chronik* appears. If we consult the topics list, we see that they share only one topic, "küning, sas, sancti, ulrici, vita, herren". While "sancti", "ulrici" and "vita" only appear in the *Leben*, "küning", "sas" and "herren" only show up in the *Chronik*. The shared topic explains the visible connection, but apparently there is no textual/content based relation between the two documents. It seems that caution is advised whenever two documents are only connected through one single topic.

the isolation of the two edited chronicles of Richental and Etterlin as well as of the codex Heidelberg 116. In the document-topic-matrix, each of these has one very large topic and only very few shared with others. Codex Heidelberg, Cpg 116 is composed of chapter 6 of the Twinger chronicle, the *Weißenburger Chronik* by Eikhart Artzt, and the *Trotula* – texts that are unique in the corpus. That this results in them being outliers is very well demonstrated in the network.

So how can these findings be interpreted? The second result seemed promising, hinting towards the relation of manuscripts that contain the same text(s), or texts treating the same subject(s): they share the *Konstanzer Jahrgeschichten* or chronicles about the city of Constance. But this thematic focus is all but obvious from the actual words that make up the connecting topics: for the three manuscripts Freiburg im Breisgau, UB, Hs. 471, Munich, BSB, Cgm 568 and Heidelberg, UB, Cpg 475, the binding topic consists of "gottes, rome, geburt", and for Freiburg im Breisgau, UB, Hs. 471, Munich, BSB, Cgm 568 and Stuttgart, LB, HB V 22, the topic is made up of "kùng, kayser, volk", which does not in any way point towards the city of Constance. For the connecting topics of the codices Wolfenbüttel, HAB, Cod. 16.17 and Dresden, UB, Mscr. F 98, we can observe something similar: The topics "stat, gottes, geburt" and "herren, lant, starp" do not contain an easy-to-read hint at the connecting texts and/or topics in the particular manuscript. One explanation for the composition of the topics might be found in medieval writing practices: events are often dated referring to the birth of Christ, the ruling emperor, or the current pope, and as most of the manuscripts in this corpus contain historiography, there are many events that are contextualized with such a reference. One could exclude words like "gottes", "geburt", and all the different forms for king and pope, using the stop word list – but this would account for a bias that might already be too large, by eliminating words in order to get to the "real" meaningful terms – and eventually concepts.[48]

The topics generated here cannot serve as designators of manuscripts in a corpus with regards to their specific content. However, they do provide additional value in tracing the relationships between manuscripts and discovering networks: during copying processes, the linguistic peculiarities of the copied manuscript are often kept within direct adoptions of (parts of) the texts, leading to a fair consistency of spelling. Without normalization or lemmatization of the texts, the resulting topics actually point towards relations between manuscripts from a linguistic point of view. If we look again at the document-topic-matrix and compare the dialects that were assigned to the two codices Wolfenbüttel, HAB, Cod. 16.17 and Dresden, UB, Mscr. F 98, which to our knowledge do not share any texts nor treat similar subjects, this assumption can be confirmed (see Fig. 11).

48 For a discussion of the influence of stop word lists, see Hodel, Möbus, and Serif, "Von Inferenzen und Differenzen.

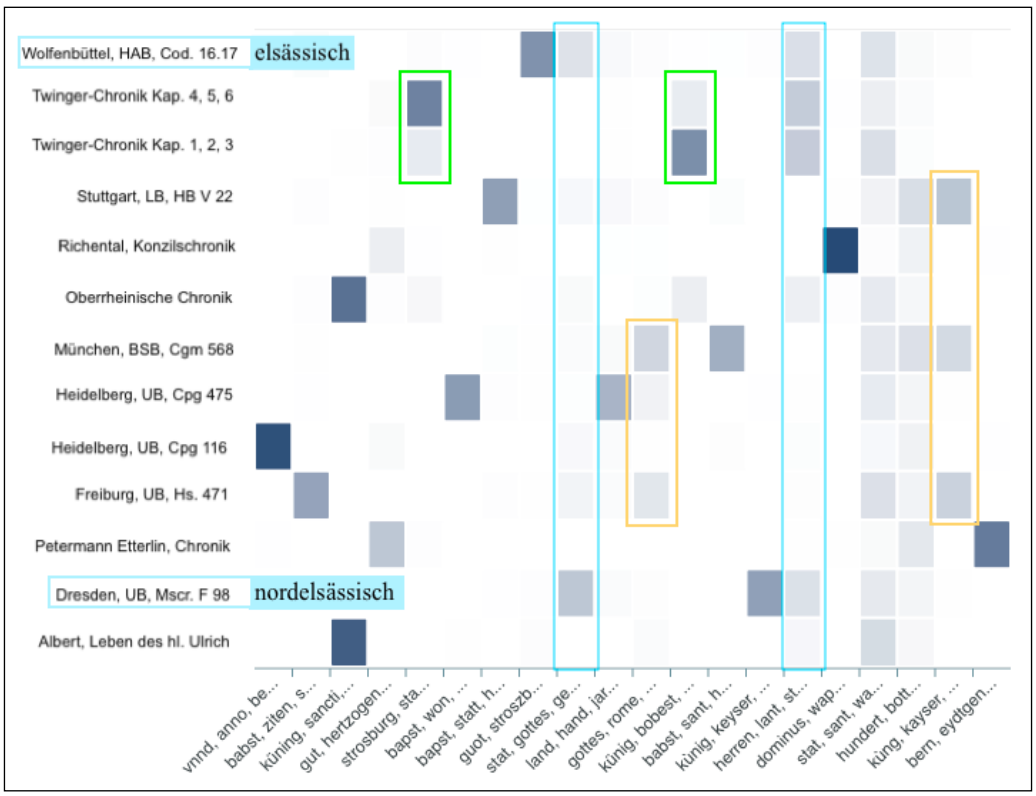

Fig. 11 Document-topic-matrix with weighed prevalence of a topic (13 documents, 20 topics), and writing dialects for the two related codices Wolfenbüttel, HAB, Cod. 16.17 and Dresden, UB, Mscr. F 98, as provided in the existing manuscript descriptions.

Both manuscripts are written in an Alsatian dialect, so the relation suggested through the topics very likely lies more in the writing of "lant" or "gottes" (instead of, e.g., "lande" or "gotz", spellings that are found in other codices in the corpus) than in the hitherto unknown shared transmission of (a) text(s). And while "elsässisch" and "nordelsässisch" are not too problematic if one considers the comparability of such metadata, things are more complicated for other codices (see Fig. 12).

While for Stuttgart, LB, HB V 22, we only know the place of production – as Constance is in an area with an Alemannian dialect – the label for the dialects of the other three codices do not seem too similar at first glance; however, they can in fact be put in close proximity (see Fig. 13).[49]

49 Based on Peter Wiesinger, "Die Einteilung der deutschen Dialekte." In *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung,* 2nd half-vol., edited by

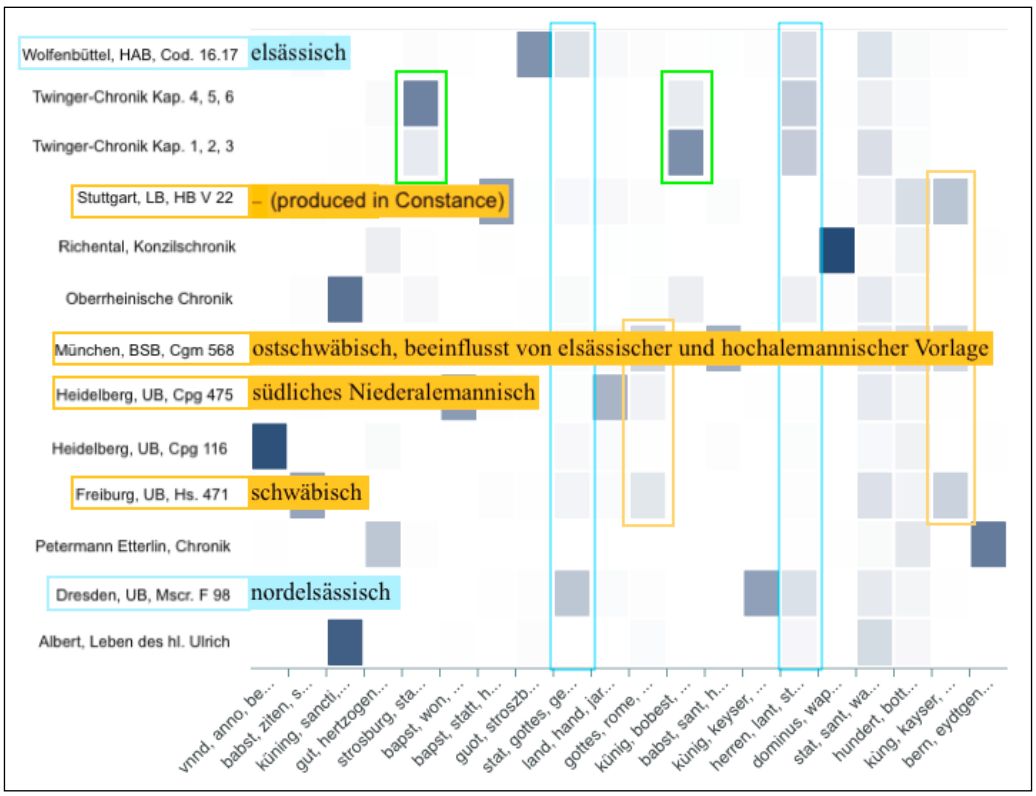

Fig. 12 Document-topic-matrix with weighed prevalence of a topic (13 documents, 20 topics), and writing dialects for the four related codices Freiburg im Breisgau, UB, Hs. 471, Heidelberg, UB, Cpg 475, Munich, BSB, Cgm 568 and Stuttgart, LB, HB V 22.

Of course, a proximity of writing dialects does not guarantee a proximity of the places of production – but it is a very good indicator for potentially related manuscripts, as is also suggested by the shared transmission of the *Konstanzer Jahrgeschichten,* the account of the murder of the bishop of Lausanne and a list of the bishops of Constance. In contrast to the information about the writing dialect in plain text, as is given in manuscript descriptions, without standardization or a universal classification system, the relation between the manuscripts through topics can be taken as additional and non-interpretive metadata, which can also be generated for codices which do not have a highly detailed description.

Werner Besch et al. Berlin/New York, 1983, 831. Even though the classification reflects the dialectal areas at the beginning of the twentieth century, the map can serve as a first indication for the localization of manuscripts.

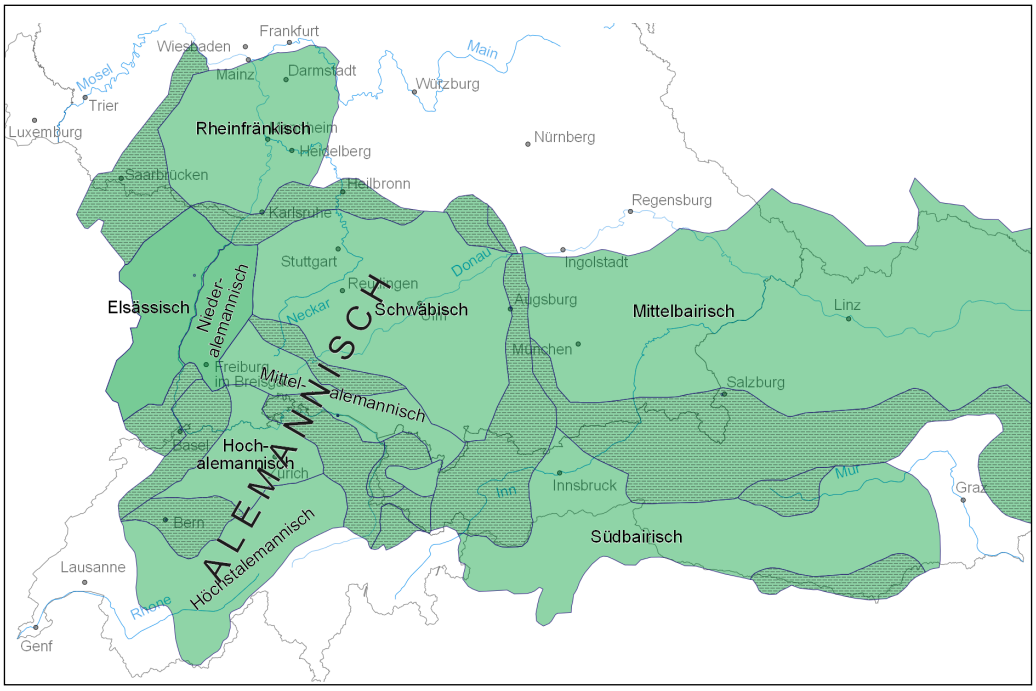

Fig. 13 Dialectal areas in southwest Germany, based on the classification by Peter Wiesinger. Own map created using SprachGIS (www.regionalsprache.de), 2023. The crosshatched areas were classified as transition zones that cannot be assigned to a specific dialect.

5. Conclusion

This article proposed the application of topic modeling to premodern manuscripts, in particular compilations, as an additional method for the classification of codices within a corpus. On the basis of this classification, the connections as well as the networks between multiple-text manuscripts can be illustrated, getting around some of the shortcomings of a purely text- or genre-based approach.

For the tracing of medieval manuscript transmission, network analysis based on single texts as connectors between codices has been successfully applied in both previous and current research. For certain corpora, this is a promising and fruitful approach that highlights the processes of copying and compiling, as is shown by the contributions in this issue. But it has its limits, among other things resulting from the imperfect metadata available for the majority of medieval manuscripts – one has to either rely on existing manuscript descriptions, or conduct a close reading of the documents in question, an undertaking which is not always feasible. To better understand the reading interests or intended functions of a specific compilation and its further circulation, a broader perspective, which

considers its complete content, could be insightful. Networks based on the genres of the distinct texts contained in a manuscript could be built and used for further exploration of a corpus. But genre-based classification always contains an interpretive element, and there is no universal ontology or taxonomy that can be used as a reference. To compensate for missing or inaccurate manuscript descriptions and to eliminate subjective or idiosyncratic genre attributions, topic modeling can be a viable method. Computationally calculated topics are inferred from the full text of a manuscript, independent of different stages of production or changing writing hands, thus taking into account the whole document.

Topic modeling calculates topics for a document in relation to the other documents in a specific corpus (which could also consist of the entirety of all existing manuscripts, of course). It therefore facilitates or enables comparison and can serve as an exploratory tool; material and paleographic analyses would need to follow, but they could benefit from pointers towards specific (groups of) manuscripts. And while some of the topics discussed in the examples above reflect dating conventions rather than revealing hidden content, it seems plausible that by curating distinctive lists of stop words and comparing their outcomes, topics could become more meaningful with respect to a thematic programme, and likely to the intention of the writer or their client – of course, attention has to be paid not to include another kind of interpretive bias through the excluded words. Developments and improvements in the field of text recognition and normalization and/or lemmatization for pre-modern languages would also add to the method, with the latter helping to focus more on content, but this seems to remain an unsolved problem for the moment.[50] A two-step approach of topic modelling, first with raw text, then with normalized text, could deliver both reliable results with regards to writing language as well as textual content of the manuscripts in a specific corpus.

At present, the apparent superficiality of the results – with regards to the "real" content – does not reduce the value added by the topics calculated. They provide useful indications towards relations between manuscripts without having to rely on mostly non-standardized metadata and without resulting in yet another tax-

50 The normalizer "Norma" is currently not being developed further, as the latest release dates from 2017; see https://www.linguistics.rub.de/comphist/resources/norma/index.html and https://github.com/comphist/norma. For a comparison of different approaches of normalization, including Norma, see Marcel Bollmann, "A Large-Scale Comparison of Historical Text Normalization Systems," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers)* (Minneapolis, MN, 2019), 3885–98, doi:10.18653/v1/N19-1389; Simon Flachs, Marcel Bollmann, and Anders Søgaard, "Historical Text Normalization with Delayed Rewards," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence: Association for Computational Linguistics, 2019), 1614–19, doi:10.18653/v1/P19-1157.

onomy of subjective classifications. Experimenting with different forms of visualizations helps to interpret the results, while heatmaps and networks complement each other. In the worst case, the calculated connections merely point towards similarities between codices on a linguistic level, or to general textual similarities such as dating conventions. With the constant improvement of handwritten text recognition, by experimenting with stop words, and considering the wait for functioning normalizers, the results can certainly be improved, showing topic-based networks between manuscripts that help to better understand their production and transmission.

6. References

Bausi, Alessandro, Michael Friedrich, and Marilena Maniaci, eds. *The Emergence of Multiple-Text Manuscripts.* Studies in Manuscript Cultures 17. Berlin/Boston, 2019.

Blei, David M. "Probabilistic Topic Models." *Communications of the ACM* 55, no. 4 (April 2012): 77–84.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." *The Journal of Machine Learning Research* 3 (March 1, 2003): 993–1022.

Bollmann, Marcel. "A Large-Scale Comparison of Historical Text Normalization Systems." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers),* 3885–98. Minneapolis, MN, 2019. doi:10.18653/v1/N19-1389.

Buck, Thomas Martin, ed. *Chronik des Konstanzer Konzils 1414–1418 von Ulrich Richental: historisch-kritische Edition. Band 1: A-Version.* Vol. XLIX, 1–3. Konstanzer Geschichts- und Rechtsquellen. Ostfildern, 2020.

Cerquiglini, Bernard. *Éloge de la variante. Histoire critique de la Philologie.* Paris, 1989.

Fernández Riva, Gustavo. "Network Analysis of Medieval Manuscript Transmission. Basic Principles and Methods." *Journal of Historical Network Research* 3 (2019): 30–49.

Flachs, Simon, Marcel Bollmann, and Anders Søgaard. "Historical Text Normalization with Delayed Rewards." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 1614–19. Florence: Association for Computational Linguistics, 2019. doi:10.18653/v1/P19-1157.

Friedrich, Michael, and Cosima Schwarke, eds. *One-Volume Libraries. Composite and Multiple-Text Manuscripts.* Studies in Manuscript Cultures 9. Berlin/Boston, 2016.

Geith, Karl-Ernst, ed. *Albert von Augsburg: Das Leben des heiligen Ulrich.* Quellen und Forschungen zur Sprach- und Kulturgeschichte der germanischen Völker, n.F. 39 (163). Berlin/New York, 1971.

Graf, Klaus. "Die verschollene Twinger-Handschrift aus dem Regularkanonikerstift Dürnstein." *Archivalia,* March 27, 2013. https://archivalia.hypotheses.org/6941.

Grieshaber, Franz Karl. *Oberrheinische Chronik: Älteste bis jetzt bekannte, in deutscher Prosa.* Rastatt, 1850.

Gruber, Eugen, ed. *Petermann Etterlin. Kronica von der loblichen Eydtgnoschaft, Jr harkommen und sust seltzam stritten und geschichten.* Quellenwerk zur Entstehung der Schweizerischen Eidgenossenschaft 3, 3. Aarau, 1965.

Hegel, Carl, ed. *Die Chroniken der oberrheinischen Städte. Straßburg 1.* Die Chroniken der deutschen Städte vom 14. bis ins 16. Jahrhundert 8. Leipzig, 1870.

— ed. *Die Chroniken der oberrheinischen Städte. Straßburg 2.* Die Chroniken der deutschen Städte vom 14. bis ins 16. Jahrhundert 9. Leipzig, 1871.

Hillenbrand, Eugen. "Gallus Öhem, Geschichtsschreiber der Abtei Reichenau und des Bistums Konstanz." In *Geschichtsschreibung und Geschichtsbewußtsein im späten Mittelalter,* edited by Hans Patze, 727–55. Vorträge und Forschungen 31. Sigmaringen, 1987.

Hodel, Tobias, Dennis Möbus, and Ina Serif. "Von Inferenzen und Differenzen. Ein Vergleich von Topic-Modeling-Engines auf Grundlage historischer Korpora." In *Von Menschen und Maschinen. Mensch-Maschine-Interaktionen in digitalen Kulturen,* edited by Selin Gerlek et al., 181–205. Hagen, 2022.

Johnston, Michael, and Michael Van Dussen. "Introduction: Manuscripts and Cultural History." In *The Medieval Manuscript Book. Cultural Approaches,* edited by Michael Johnston and Michael Van Dussen, 2–16. Cambridge Studies in Medieval Literature 94. Cambridge, 2015.

Julien, Octave. "Délier, lire et relier. L'Utilisation de l'analyse réseau pour construire une typologie de recueils manuscrits de la fin du Moyen Âge." *Hypothèses* 19, no. 1 (2016): 211–24. doi:10.3917/hyp.151.0211.

Löser, Freimut. "Überlieferungsgeschichte(n) schreiben." In *Überlieferungsgeschichte transdisziplinär: Neue Perspektiven auf ein germanistisches Forschungsparadigma,* edited by Dorothea Klein, Horst Brunner, and Freimut Löser, 1–19. Wissensliteratur im Mittelalter 52. Wiesbaden, 2016.

Maniaci, Marilena. "Miscellaneous Reflections on the Complexity of Medieval Manuscripts." In *Collecting, Organizing and Transmitting Knowledge. Miscellanies in Late Medieval Europe,* edited by Sabrina Corbellini, Giovanna Murano, and Giacomo Signore, 11–22. Bibliologia: Elementa ad Librorum Studia Pertinentia 49. Turnhout, 2018.

Mone, Franz Josef, ed. *Quellensammlung der badischen Landesgeschichte 1.* Karlsruhe, 1848.

— ed. *Quellensammlung der badischen Landesgeschichte 3.* Karlsruhe, 1863.

Müller, Diana. *Textgemeinschaften. Der "Gregorius" Hartmanns von Aue in mittelalterlichen Sammelhandschriften.* Frankfurt a.M., 2013. http://publikationen.ub.uni-frankfurt.de/frontdoor/index/index/docId/30069.

Mutuvi, Stephen, Antoine Doucet, Moses Odeo, and Adam Jatowt. "Evaluating the Impact of OCR Errors on Topic Modeling." In *Maturity and Innovation in Digital Libraries,* edited by Milena Dobreva, Annika Hinze, and Maja Žumer, 3–14. Lecture Notes in Computer Science. Cham, 2018. doi:10.1007/978-3-030-04257-8_1.

Nichols, Stephen G. "What is a Manuscript Culture? Technologies of the Manuscript Matrix." In *The Medieval Manuscript Book: Cultural Approaches,* edited by Michael Johnston and Michael Van Dussen, 34–59. Cambridge Studies in Medieval Literature 94. Cambridge, 2015.

Nichols, Stephen G., and Siegfried Wenzel, eds. "Introduction." In *The Whole Book: Cultural Perspectives on the Medieval Miscellany,* 1–6. Recentiores. Later Latin Texts and Contexts. Michigan, 1996.

Pielström, Steffen, Severin Simmler, and Thorsten Vitt. "Topic Modeling with Interactive Visualizations in a GUI Tool." *Proceedings of the Digital Humanities Conference Utrecht 2019,* n.d. https://dev.clariah.nl/files/dh2019/boa/0637.html.

Purschwitz, Anne. "Netzwerke des Wissens – thematische und personelle Relationen innerhalb der Halleschen Zeitungen und Zeitschriften der Aufklärungsepoche (1688–1818)." *Journal of Historical Network Research* 2 (December 3, 2018): 109–42.

Robertson, Stephen. "Understanding Inverse Document Frequency: On Theoretical Arguments for IDF." *Journal of Documentation* 60, no. 5 (January 1, 2004): 503–20. doi:10.1108/00220410410560582.

Scherrer, Gustav. *Kleine Toggenburger Chroniken. Mit Beilagen und Erörterungen.* St. Gallen, 1874.

Schneider, Karin. "Berufs- und Amateurschreiber. Zum Laien-Schreibbetrieb im spätmittelalterlichen Augsburg." In *Literarisches Leben in Augsburg während des 15. Jahrhunderts,* edited by Johannes Janota and Werner Williams-Krapp, 8–26. Studia Augustana 7. Tübingen, 1995.

Schöch, Christof. "Computational Genre Analysis." In *Digital Humanities for Literary Studies: Methods, Tools, and Practices,* edited by James O'Sullivan, 219–31. College Station, TX, 2020.

— "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama." *Digital Humanities Quarterly* 11, no. 2 (2017): §1–53. https://doi.org/10.48550/arXiv.2103.13019.

Schuhmacher, Mareike. "DARIAH Topics Explorer." *ForTEXT. Literatur digital erforschen.* Accessed October 29, 2021. https://fortext.net/tools/tools/dariah-topics-explorer.

Serif, Ina. "Der zerstreute Chronist. Zur Überlieferung der deutschsprachigen Chronik Jakob Twingers von Königshofen." *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte,* May 12, 2015. https://mittelalter.hypotheses.org/7063, last updated June 28, 2023.

— *Geschichte aus der Stadt. Überlieferung und Aneignungsformen der deutschen Chronik Jakob Twingers von Königshofen.* Kulturtopographie des alemannischen Raums 11. Berlin/Boston, 2020.

— "Konstanzer Jahrgeschichten." In *Encyclopedia of the Medieval Chronicle, 2nd Online Edition,* edited by Graeme Dunphy and Christian Bratu. Leiden/Boston, 2016. http://dx.doi.org/10.1163/2213-2139_emc_SIM_001450.

Snijders, Tjamke. "Work, Version, Text and Scriptum: High Medieval Manuscript Terminology in the Aftermath of the New Philology." *Digital Philology. A Journal of Medieval Cultures* 2, no. 2 (2013): 266–96.

Stolz, Michael. "Linking the Variance. Unrooted Trees and Networks." In *The Evolution of Texts. Confronting Stemmatological and Genetical Methods. Proceedings of the International Workshop Held in Louvain-La Neuve on September 1–2, 2004,* edited by Caroline Macé, 193–213. Linguistica Computazionale 24/25. Pisa, 2006.

Underwood, Ted. "Topic modeling made just simple enough," July 4, 2012. https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/.

Weingart, Scott B. "Topic Modeling for Humanists: A Guided Tour," July 25, 2012. http://scottbot.net/topic-modeling-for-humanists-a-guided-tour/.

— "Topic Nets," November 10, 2012. http://scottbot.net/topic-nets/.

Wiesinger, Peter. "Die Einteilung der deutschen Dialekte." In *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung,* 2nd half-vol., edited by Werner Besch et al. Berlin/New York, 1983, 807–900.

Westphal-Wihl, Sarah. *Textual Poetics of German Manuscripts, 1300–1500*. Studies in German Literature, Linguistics, and Culture. Columbia, SC, 1993.

Wolf, Jürgen. "Sammelhandschriften – mehr als die Summe ihrer Einzelteile." In *Überlieferungsgeschichte transdisziplinär: Neue Perspektiven auf ein germanistisches Forschungsparadigma,* edited by Dorothea Klein, Horst Brunner, and Freimut Löser, 69–81. Wissensliteratur im Mittelalter 52. Wiesbaden, 2016.