Acta Historica Tallinnensia 32/2 https://doi.org/10.3176/hist.2026.2.02

The Historian and Modern Statistical Methods: Applications to Thirteenth-century Livonian Chronicles

Nicolas Bourgeois
Urban Radar, Fougères, France; nbourgeo@phare.normalesup.org

Abstract. The article examines the application of statistical and machine-learning-based tools to two thirteenth-century Livonian chronicles: the *Chronicle of Henry of Livonia* and the *Livonian Rhymed Chronicle*. It highlights both the potential and the limitations of these methods for the analysis of medieval texts. The article discusses key challenges, including sparse and multilingual corpora, difficulties of evaluation, and methodological complexity. The article argues that the principal value of increasingly complex computational methods lies less in autonomous historical interpretation than in their integration into transparent, verifiable, and collaboratively curated scholarly infrastructures. It therefore proposes principles for designing digital tools and metadata that are accessible, reliable, and open to continuous scholarly verification.

Keywords: medieval chronicles, medieval Livonia, lexicometry, clustering, topic modelling, embeddings, transformers

The pace of progress in statistical methods, particularly in the field of machine learning, often rebranded as artificial intelligence, has created a gap between the capabilities of these systems and researchers' ability to keep pace with their development, even when those researchers have a technical background. Many open systems are also presented through technical language, abstract concepts, or opaque interfaces that make

© 2026 Author. This is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 International License CC BY 4.0 (http://creativecommons.org/licenses/by/4.0).

it difficult even for scholars with high levels of digital literacy to grasp their full potential – and their serious limitations. On the other hand, some tools now available to a general audience are remarkably easy to use, requiring little or no knowledge of their underlying mechanisms, while also producing highly convincing results. This combination creates both an urgent need for validation and a significant challenge, not only to determine how such outputs should be validated, but also to acquire and master the techniques required to do so effectively. As a result, researchers may underestimate what is possible, feel overwhelmed and avoid engaging, or, conversely, display excessive confidence in results they may not have the capacity to assess.

This article approaches these problems through two thirteenth-century Livonian chronicles: the *Chronicle of Henry of Livonia* and the *Livonian Rhymed Chronicle*. They provide useful test cases because they combine several difficulties characteristic of medieval textual corpora: limited size, different languages, unstable historical vocabularies, complex textual transmission, and substantial differences of genre, authorship, and communicative context. Rather than asking whether computational methods can replace close reading or historical interpretation, we ask more specifically: What can different levels of computational text analysis contribute to the study of medieval historical texts? Where do their evidential limits become apparent? And under what conditions can increasingly complex and opaque models be integrated responsibly into historical research?

Our central argument is that the principal value of such methods lies less in producing autonomous historical interpretations than in supporting the construction of transparent, verifiable, and collectively curated scholarly resources. Simpler techniques such as lexicometry, clustering, and topic modelling can provide useful exploratory perspectives when their assumptions remain visible to the researcher. More complex methods, including embeddings and large language models, may be harder to interpret directly, but they can still be valuable when used to generate, organise, or enrich metadata within controlled scholarly workflows. In this sense, computational models are most productive when understood not as substitutes for historical interpretation, but as part of a broader paratextual infrastructure comparable to indexes, regesta, critical apparatuses, and scholarly editions.

The article develops this argument in several stages. We begin with comparatively transparent methods based on word frequency, clustering, and dimensionality reduction, examining both their usefulness and their

limitations when applied to the two Livonian chronicles. We then turn to probabilistic topic modelling and embeddings, which allow more complex relationships between words and passages to be explored but also introduce greater interpretative and methodological difficulty. Finally, we consider the use of transformer-based language models for metadata generation and argue that their most promising role in historical research lies in collaboratively designed and carefully documented systems in which computational outputs remain open to scholarly verification and revision.

NATURAL LANGUAGE PROCESSING FOR TEXT ANALYSIS

Text analysis has evolved considerably over the past sixty years, as computer scientists have taken advantage of increasing computing resources to design algorithms that can treat texts less as bags of isolated words and more as complex structures in which each word derives its full meaning from its relation to others.[1]

The early stages of this evolution focused on incorporating the immediate context within a given text, from simple counts of co-occurrences to more in-depth analyses of grammatical structure. These methods have long been integrated into the toolkits used by historians familiar with quantitative methods.[2] In addition, statisticians and machine learning specialists have applied techniques such as supervised classification, clustering, regression, and dimensionality reduction to generate insights, metrics, and visualisations from large corpora of texts.[3]

By contrast, recent breakthroughs have focused on building statistical models of language itself, validated by their ability to perform some relatively easy-to-evaluate tasks such as sentence completion, rather than by any a priori understanding of linguistic structure. These advances were initially driven by broader tasks such as translation,

1 M. Piotrowski. Natural Language Processing for Historical Texts. (Synthesis Lectures on Human Language Technologies, 5). Springer, Cham, 2012. https://doi.org/10.1007/978-3-031-02146-6.

2 N. Bourgeois, M. Cottrell, S. Lamassé, M. Olteanu. Search for Meaning Through the Study of Co-occurrences in Texts. – Advances in Computational Intelligence. IWANN 2015. Lecture Notes in Computer Science, 9095. Ed. by I. Rojas, G. Joya, A. Catala. Springer, Cham, 2015, 578–591. https://doi.org/10.1007/978-3-319-19222-2_48.

3 N. Bourgeois, A. Pellet, M. Puren. Using Topic Generation Model to Explore the French Parliamentary Debates During the Early Third Republic (1881–1899). – Digital Humanities in the Nordic and Baltic Countries Publications, 2022, 4, 2, 35–51. https://doi.org/10.5617/dhnbpub.11259.

question answering, and image captioning, but were quickly applied to text analysis. The resulting tools, including recurrent neural networks and transformers,[4] have proven extremely powerful, but also increasingly difficult to interpret and evaluate.

As a result, historians now have access to a wide range of text analysis methods that are increasingly sophisticated, but also opaque and difficult to master. We argue that all of these techniques can be useful for researchers, though not in the same way. Statisticians should design tools that help historians make use of the most recent methods not only by making them easy to use, but, more importantly, by ensuring that they are easy to understand, evaluate, and safely integrate into scholarly workflows.

On the other hand, historians could benefit from developing a solid familiarity with earlier, more established techniques, since mastery of these provides the best safeguards for engaging with more recent approaches. We therefore advocate incorporating these technologies into a collective process aimed at building, proofreading, sharing, curating, and ultimately using secondary materials, rather than using them in stand-alone studies that are difficult for a wider scholarly audience to understand or replicate.

WRITTEN SOURCES FOR THIRTEENTH-CENTURY LIVONIA

In this article, we set aside non-literary evidence, such as archaeological or numismatic finds. This in no way means that such evidence cannot provide valuable material for automated analysis. This is the case not only because it often contains inscriptions, but also because image analysis can be fully integrated into language analysis. For instance, recent transformer-based vision–language models construct a latent space in which both text and images are jointly encoded. To give an intuitive example, an image of a sword would be assigned a representation close to

4 S. Hochreiter, J. Schmidhuber. Long Short-term Memory. – Neural Computation, 1997, 9, 8, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735; A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin. Attention is All You Need. – Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, 6000–6010. https://dl.acm.org/doi/10.5555/3295222.3295349; J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. – 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). Association for Computational Linguistics, Minneapolis, 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423.

that of the word 'sword' itself.[5] However, incorporating these methods would add a significant layer of complexity to an article that is primarily intended as an introduction to the field.

With that restriction in mind, the documentary basis for studying thirteenth-century Livonia falls broadly into two categories: narrative accounts and normative documents. Some have been preserved in their original form, while others have come down to us only through later manuscript traditions and scholarly compilations, which may have introduced changes in vocabulary.

The most structurally informative sources are legal and administrative documents. These include charters, privileges, treaties, monastic or episcopal statutes, and, above all, an impressive corpus of papal letters, whose numerical prominence is undoubtedly related to the better preservation conditions of pontifical archives. Together, these sources form the backbone of our understanding of the early history of Livonia.[6] Most are written in Latin, although a significant number of German texts also survive. These documents vary considerably in their form, vocabulary, and scope, and thus constitute the closest, albeit much scarcer, equivalent to the corpora used to train large language models. For the same reason that we do not use archaeological sources, we do not include these texts in this preliminary study. Nevertheless, they should be regarded as valuable material for any serious quantitative text analysis and as absolutely necessary for any attempt to train a model on the language used in the chronicles.

This leaves us with narrative sources only. Among them, the *Chronicle of Henry of Livonia* remains foundational. Written in Latin in the 1220s by the priest Henry, it provides a detailed, near-contemporary account of missionary activity, conquest, and the establishment of ecclesiastical structures, as well as extensive information about the eastern Baltic societies of the time. It is undoubtedly the most detailed and best-studied text on medieval Livonia and, for these reasons, will also be the main focus of our study.[7]

5 A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever. Learning Transferable Visual Models from Natural Language Supervision. – Proceedings of the 38th International Conference on Machine Learning. Vol. 139 of Proceedings of Machine Learning Research, 2021, 8748–8763. https://doi.org/10.48550/arXiv.2103.00020.

6 Liv-, est- und kurländisches Urkundenbuch nebst Regesten. Erster Band 1093–1300. Hrsg. von F. G. von Bunge. Kluge und Ströhm, Reval, 1853.

7 See Crusading and Chronicle Writing on the Medieval Baltic Frontier. A Companion to the Chronicle of Henry of Livonia. Ed. by M. Tamm, L. Kaljundi, C. S. Jensen. Ashgate, Farnham, 2011.

A complementary perspective appears in the *Livonian Rhymed Chronicle*, composed later in the thirteenth century in vernacular Middle High German poetic form. This text reflects the viewpoint of the successive military orders – the Brothers of the Sword and the Livonian branch of the Teutonic Order – and places greater emphasis on military events. Its narrative priorities and chronological precision differ considerably from those of Henry's chronicle, and individual episodes cannot always be reconciled with other surviving evidence. It is important to include it in our study because, while referring to many of the same or similar events, it employs a markedly different vocabulary, shaped not only by its vernacular language but also by its intended audience, purpose, and date of composition.[8]

For the purpose of this exploratory article, we use digital editions of the chronicles, which are based on the automated optical character recognition (OCR) of Leonid Arbusow's and Albert Bauer's 1955 edition of Henry's chronicle, referred to as HCL,[9] and Leo Meyer's 1876 edition of the *Livonian Rhymed Chronicle*, referred to as RC.[10] In the case of the RC, while the OCR quality is sufficient for a methodological introduction, any attempt to build reliable metadata should, of course, rely instead on a proofread digital edition, including an additional layer of linguistic disambiguation.

In the first part of this article, we focus on the possibilities and limitations of traditional lexicometric methods, including clustering and dimensionality reduction. These techniques can be used without an extensive mathematical background. In the second and third parts, we introduce more sophisticated tools based on robust statistical models, including latent Dirichlet allocation, latent space models, and transformers, which may be difficult for scholars to apply independently with a high degree of confidence. In the final part, we explore how the automated construction of metadata in a controlled context may help bridge this gap.

8 See A. V. Murray. The Structure, Genre and Intended Audience of the Livonian Rhymed Chronicle. – Crusade and Conversion on the Baltic Frontier, 1150–1500. Ed. by A. V. Murray. Ashgate, Aldershot, 2001, 235–251; A. V. Murray. Formulaic Language in the *Livonian Rhymed Chronicle*: Set Phrases and Discourse Markers in Middle High German History Writing. – Amsterdamer Beiträge zur älteren Germanistik, 2019, 79, 86–105; A. V. Murray. The *Livonian Rhymed Chronicle* as a Transitional Text: Formulaic Language in Middle High German Verse History. – Rilce: Revista de Filología Hispánica, 2020, 36, 4, 1324–1343. https://doi.org/10.15581/008.36.4.1324-43.

9 Heinrichs Livländische Chronik = Heinrici chronicon Livoniae. Bearb. von L. Arbusow, A. Bauer. Hahnsche Buchhandlung, Hannover, 1955.

10 Livländische Reimchronik, mit Anmerkungen, Namenverzeichnis und Glossar. Hrsg. von L. Meyer. F. Schöningh, Paderborn, 1876.

FROM WORD FREQUENCY TO CONTEXT ANALYSIS

At the most basic level of automated analysis lies lexicometry, an approach that treats a text as a collection of measurable units. In this framework, words are counted, compared, and statistically analysed without regard to grammatical structures or more sophisticated local statistical patterns, although the approach still has many advantages. It is easy to understand with a minimal mathematical background, and the errors it produces due to its rudimentary nature (homonyms, non-standard spelling, lack of contextual information) are equally easy to interpret, if not to prevent. It can be very useful at the exploratory stage for gaining insight into a large corpus, or at a more advanced stage to test a given hypothesis.

The combination of lexicometry with close reading and traditional textual analysis provides a compelling example of how historians can incorporate quantitative methods into the construction of an argument. Wojtek Jezierski has previously examined the vocabulary of emotions across the two chronicles, using word frequencies both to illustrate and to support his interpretation.[11] Rather than relying solely on raw counts, he carefully contextualises the data, thereby avoiding the common pitfall of drawing conclusions from frequency alone. Notably, his article also suggests that topic modelling could extend this line of analysis even further.

Table 1. Most frequent words in the two chronicles, excluding stopwords: HCL before lemmatisation (top), HCL after lemmatisation (middle), and RC without lemmatisation (bottom).

castrum	castro	episcopus	lyvones	terram	milicie	exercitum
236	146	140	133	131	130	122
castrum	episcopus	exercitus	frater	terra	christianus	magnus
451	404	296	240	207	162	161
meister	brûdere	komen	sprechen	heide	manche	landen
478	321	205	205	204	203	197

11 W. Jezierski. Politics of Emotions and Empathy Walls in Thirteenth Century Livonia. – Making Livonia: Actors and Networks in the Medieval and Early Modern Baltic Sea Region. Ed. by A. Mänd, M. Tamm. Routledge, London, New York, 2020, 113–142. See also W. Jezierski. Emotional Bonding and Trust during Sieges, Twelfth–Thirteenth Centuries. – Risk, Emotions, and Hospitality in the Christianization of the Baltic Rim, 1000–1300. Brepols, Turnhout, 2022, 175–205, esp. 192–203.

If we perform a raw count of words in the HCL, we quickly encounter a well-known problem: different conjugated or declined forms of a single root are counted separately. This can be addressed relatively easily through lemmatisation, that is, by grouping together the different inflected forms of a given word, as long as the spelling remains consistent (cf. Table 1). However, lemmatisation has many shortcomings: it handles variations in spelling poorly, struggles with words that are not included in dictionaries such as *lyvones*, and performs less well with computationally under-resourced historical languages, including Middle High German. More importantly, it does not group synonyms together or distinguish between homonyms. For instance, the words *dominus* and *frater* can have different meanings throughout Henry's chronicle, as becomes evident from the following:

> *Factum est autem eodem tempore, ut augeret Dominus de die in diem numerum et familiam fratrum milicie.* (HCL XI.3, pp. 48–49)
>
> It happened also at the same time that the Lord increased from day to day the number and retinue of the Brothers of the Militia.[12]

> *Signatis itaque per Theuthoniam multis hominibus signo crucis tandem domnus episcopus ad naves revertitur, assumpto secum fratre suo Rothmaro de claustro Sygeberge.* (HCL IX.6, p. 28)
>
> Having signed many men throughout Germany with the sign of the cross, at length the bishop returned to the ships, taking with him his brother Rothmar from the monastery of Segeberg.

TOPICS BASED ON FREQUENCY

However, with a chronicle such as Henry's, written over a short period by a single author, lexicometry can be useful for comparing different parts of the text. While reliance on a single word may be risky, statistics based on frequently co-occurring groups of words may prove more robust; the combination of several elements among *frater*, *miles*, and *magister* leaves little doubt, as does the association of *mare*, *navis*, and *tempestas*. Figure 1 shows that the prevalence of one topic or another is a more reliable indicator of where the action takes place than the frequency of a single word.

12 This and all subsequent translations from Henry's chronicle are taken from *The Chronicle of Henry of Livonia*. Trans. by J. A. Brundage. The University of Wisconsin Press, Madison, 1961.

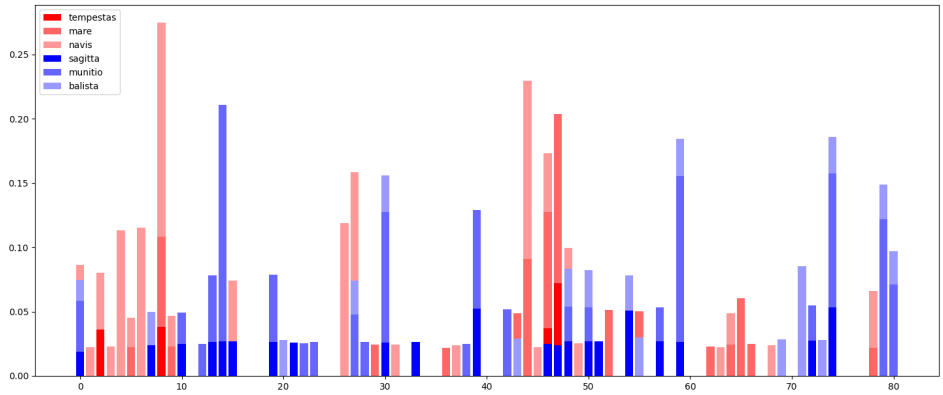


Figure 1. Comparative frequency of some words from the navigation topic (red) and from the missiles topic (blue), across consecutive two-page segments.

These groups of words, which we call topics, can be constructed automatically, without sophisticated mathematical tools, by clustering words based on their relative frequencies. Clustering is a generic technique used in data analysis to group similar data points together based on shared characteristics or patterns. In other words, we group together words that exhibit similar patterns of occurrence across the pages. There are numerous clustering techniques, such as *k*-means and agglomerative clustering, each of which has its own advantages and disadvantages depending on the structure of the data.[13]

Here, we apply a *k*-means algorithm, which groups together words with similar patterns of occurrence. For instance, according to Table 2, *balista* and *sagitta* are likely to be assigned to the same cluster, while *baptizo* will belong to a different one. For the sake of readability, we display raw counts rather than relative frequencies. However, in the computation we use term frequency–inverse document frequency (TF-IDF). TF-IDF is computed by multiplying a word's term frequency in a document by its inverse document frequency, $log(\frac{N}{df})$, where N is the total number of documents and df is the number of documents containing the word. As a result, common words receive lower weights, while distinctive words receive greater importance.

13 The scikit-learn documentation provides a useful overview of these various algorithms: https://scikit-learn.org/stable/modules/clustering.html (accessed 5/10/2026).

Table 2. Raw occurrence counts of selected words across text blocks 49–64.

block	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64
balista	0	1	0	0	0	1	1	0	0	0	1	0	0	0	0	0
sagitta	0	1	1	0	0	2	0	0	1	0	1	0	0	0	0	0
baptizo	0	1	0	0	0	0	0	3	0	1	0	0	3	3	7	2

Note that choosing a clustering method based on an arbitrary division of the text means that one occurrence of the first word in block 55 versus two occurrences of the second word in block 54 may contribute to an increase in the distance between the two words, even though they refer to the same event. In fact, the two occurrences of *balista* are in the same paragraph:

> *De obsidione castri Wendorum. Descendebant quoque sagittarii fratrum milicie de castro suo, intrabant ad Wendos et* ***balistis*** *suis multos Ruthenorum interficiebant et quam plurimos vulnerabant. (...) Mane quoque facto rex Nogardie videns multorum nobilium suorum lesionem et aliorum interfectionem castrumque ipsum Wendorum se capere non posse considerans, cum sit tamen minus castellum, quod habet Lyvonia, pacifice loquutus est fratribus milicie. Qui pacem talem non curantes* ***balistis*** *eos a se repulerunt.* (HCL XXII.5, pp. 151–152)
>
> The archers of the Brothers of the Militia came down from their fort, got through to the Wends, and with their ballistas killed many Russians and wounded a great many of them (...). When it was morning the king of Novgorod, seeing that many of his nobles were wounded and others were killed, and considering that he could not take the fort of the Wends, though, indeed, it was one of Livonia's lesser castles, spoke peacefully to the Brothers of the Militia. They cared nothing for such a peace and drove the Russians away with ballistas.

In this case, a straightforward solution would be to use Henry's textual structure instead of page numbers, since the former is closer to a semantic division than the latter. However, this would not solve another problem, which is that successive chapters would not be considered closer to each other than any other pair. A more effective approach is to use a progressive distance between word occurrences. Instead of defining $d(x, y) = 0$ if x and y belong to the same entity (page, paragraph, etc.) and 1 otherwise, we can use something like $d(x, y) = \log(1 + n)$, where n is the number of words, sentences, paragraphs, or pages between the two occurrences. The choice of a logarithmic function reflects the idea that when two words

are very far apart, the exact distance matters little, whereas variations in close proximity are more significant.

In Table 3, we display the first six words of selected topics identified by the algorithm. We observe a clear proximity among words within each topic. Topic 8, for instance, is strongly centred around the baptism process, while topic 7 is clearly more concerned with military action. A large topic, topic 6, contains the vocabulary that is distributed fairly regularly throughout Henry's chronicle. This is an artefact of the method: since a word cannot belong to several topics at once, a very common word like *exercitus* or *castrum* cannot be attached to less common words from the same lexical field that follow a more irregular pattern. For the same reason, all topics are marginally present on every page, because some of their more common words might appear here and there, which makes the results quite noisy.

Table 3. Examples of topics identified by k-means over TF-IDF (first words, n=20). Topic 6 has been omitted since it contained more than half of the vocabulary and was the 'default' topic, while topic 2 only contained three very common words (omnis, qui, sum).

1	acceptus	captivans	circuiverunt	consequor	consolatio	fraternitas
3	apostolicus	audeo	benedicens	benigne	celestia	concilium
4	capio	congrego	duco	evado	fluvius	fuga
5	abhorrens	abicio	adaptatus	affigo	affirmo	antiquus
7	adoro	advocacias	affligens	affligo	andreas	animadverto
8	accipio	baptismum	beatus	captivus	christus	dominus

Table 4. The beginning of Henry's chronicle, with each word mapped to a topic. Some words (for example *ydolatrie*) were not lemmatised because they are absent from the dictionary and different forms of the same word may therefore be assigned to different topics. A deeper study of the chronicle could include pre-lemmatisation orthographic normalisation and/or adding text-specific entries to the dictionary.

(divina, 6) (providencia, 6) (memor, 5) (raab, 6) (babilonis, 6) (videlicet, 9) (confusus, 15) (gentilitas, 6) (nostris, 6) (modernus, 6) (tempus, 10) (volens, 9) (modernus, 6) (livones, 6) (ydolatras, 6) (ydolatrie, 14) (peccatum, 6) (sompno, 6) (taliter, 1) (ignis, 6) (amor, 6) (excitavit, 6) . . .

This method tends to place extra emphasis on specificity, which may introduce artefacts. Typically, if a group of words appears several times in one location and nowhere else, it is likely to form a cluster. For instance,

as shown in Figure 2 (upper graph), topics 1 and 3 consist almost entirely of the words found on just a couple of pages (respectively HCL XX.6, p. 138 and HCL XXIX.8, p. 214).

Even these poorly clustered topics can still offer useful clues. For instance, topic 16 (*archiepiscopus*, *aspergo*, *dacie*, *dania*, *imperium*, *osiliam*, *vironie*) maps onto specific parts of Henry's chronicle: mainly chapters XXIII and XXIV, which are indeed concerned with the conflicts between the kingdom of Denmark and the Swordbrothers over the conquest of Estonia (see Figure 2, lower graph).

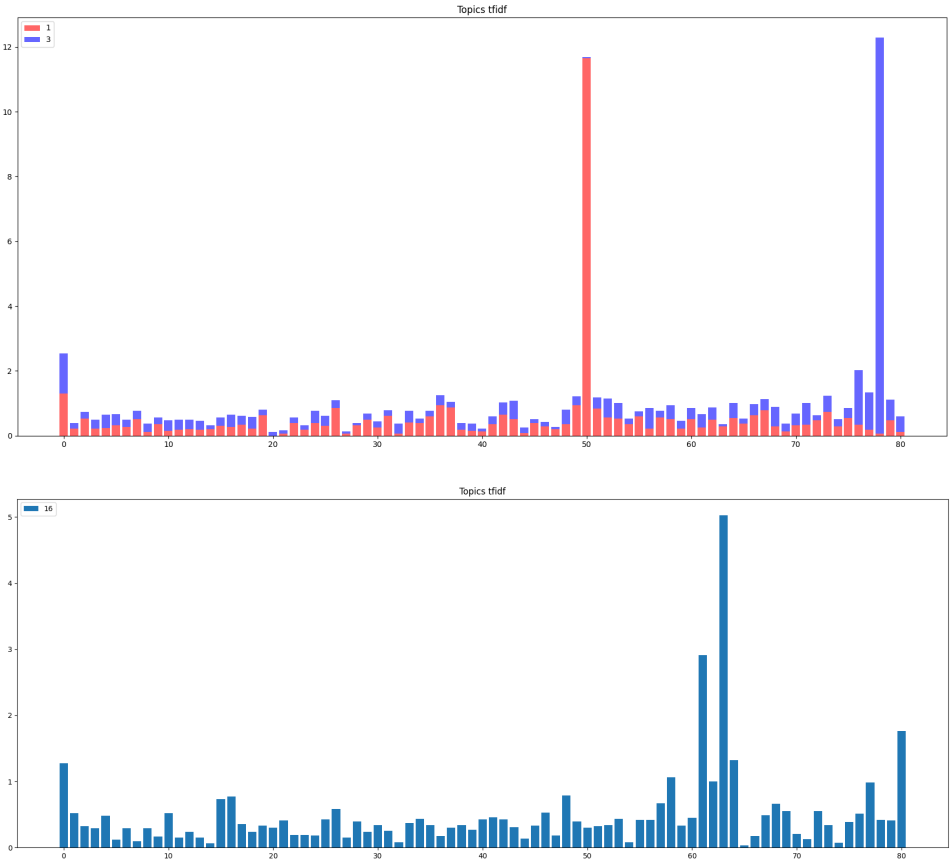


Figure 2. Comparative frequency of some TF-IDF topics across consecutive two-page segments.

CROSS-TEXT VOCABULARY

Lexicometry does not readily allow for cross-language analysis either. Certainly, in most cases, *frater* and *paganus* have the same meaning in Henry's chronicle as *brûdere* and *heide* in the *Reimchronik* – except when *frater* is used in the familial sense, for instance. In any case, direct lexicometric comparison is possible through word-for-word translation, which is often misleading. Moreover, while someone who reads both Latin and German may recognise that the following passages tell the same story, they frame it so differently that a word-count approach would not detect the correspondence.

> mit urloube daȝ geschach,
> bie der Dune ûf einen berc
> dâ bûweten sie ein êrlich werc,
> eine burc sô veste,
> daȝ die selben geste
> mit vride wol dar ûffe bliben
> und iren kouf lange triben.
> Ickesculle wart iȝ genant (...)
> Nû was als ich hân vernomen
> ein wîser man mit in komen,
> der in sanc und las,
> wen er ein reiner prîster was;
> der hêre hieȝ Meynhart. (RC, ll. 216–223, 229–233, p. 6)
>
> With native permission they built a worthy place on a mountain by the Dvina, a castle so large that these traders could remain there in peace and conduct trade for protracted periods of time. It was named Uexküll (...). Now, as I have heard it told, there had come with them a man who read and chanted the Mass for them. He was a holy priest, a man named Meinhard.

> *Fuit vir vite venerabilis et venerande caniciei, sacerdos ex ordine beati Augustini, in cenobio Sigebergensi (...) Accepta itaque licencia prefatus sacerdos a rege Woldemaro de Ploceke, cui Lyvones adhuc pagani tributa solvebant, simul et ab eo muneribus receptis, audacter divinum opus aggreditur, Lyvonibus predicando et ecclesiam in villa Ykescola construendo.* (HCL I.2–3, p. 2)
>
> In the monastery of Segeberg there was a man of worthy life, and with venerable gray hair, Meinhard by name, a priest of the order of Saint Augustine (...). After receiving, therefore, the permission of King Vladimir of Polozk, to whom the Livonians, while still pagan, paid tribute, and, at the same time, after receiving gifts from him, this priest boldly set out upon the divine work, preaching to the Livonians and building a church in the village of Uexküll.

Even texts written in the same language pose difficulties when they come from different social contexts or are written by different authors.

Clustering based on frequency is generally more effective at classifying authors than at classifying the events being described. In the next sections, we introduce two methods that can partially address these issues. Latent Dirichlet Allocation is particularly useful for identifying thematic structures across texts that share a language or sufficiently comparable vocabularies. Embeddings, by contrast, represent words or passages through their distributional contexts rather than through lexical identity alone. This makes them potentially useful for cross-linguistic comparison, but only when the representations from different languages are explicitly aligned or generated within a shared multilingual model.

LATENT DIRICHLET ALLOCATION

We have seen that the process of creating topics through clustering words based on TF-IDF variables has several shortcomings. One of them is that, by definition, clustering assigns a given word systematically to a single topic, even though it may belong to different lexical fields depending on the context (think again of *frater* and *dominus*). Another limitation is that this approach tends to favour monolithic blocks of text and struggles to account for the fact that several topics can coexist within the same statistical unit, such as a page or paragraph, depending on how the variables are constructed.

The Latent Dirichlet Allocation (LDA) topic model was first introduced in 2003.[14] It is a probabilistic generative model that represents each document as a mixture of latent topics, and each topic as a probability distribution over words. Here, a 'topic' denotes a latent statistical pattern inferred from the distribution of words across a corpus, rather than a predefined category or necessarily a theme consciously recognised by the author or reader.[15] In simplified terms, the model treats an observed text as if it had been generated by repeatedly selecting a topic from the document's topic distribution and then selecting a word from that topic's word distribution. A passage describing an expedition against the Lithuanians may therefore receive high probabilities for topics characterised by vocabulary relating to warfare and diplomacy, while receiving lower probabilities for topics dominated by seafaring

14 D. Blei, A. Ng, M. Jordan. Latent Dirichlet Allocation. – Journal of Machine Learning Research, 2003, 3, 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

15 This should come as no surprise to historians, who are accustomed to developing concepts that make sense for studying a given period, even though the people living at the time would certainly never have used them themselves – feudalism, for instance.

or ecclesiastical terminology. The analytical task is therefore to infer, from the observed distribution of words, both the word distributions associated with the latent topics and the mixture of topics associated with each textual unit. Since exact inference is computationally intractable, these distributions are estimated using approximate inference methods; here we use the variational approach introduced by Blei et al.

Latent Dirichlet Allocation has been widely used in many areas of the humanities and social sciences, as it is a very powerful tool for extracting information from a large corpus in an unsupervised context, that is, when classes are not defined a priori. In practice, topic models tend to produce more readily interpretable results when the corpus contains a sufficiently large number of textual units, when individual units concentrate on a limited range of topics, when topics recur across multiple units, and when there is enough shared vocabulary across the corpus for recurring patterns to emerge. Newspaper collections are a common example of such a corpus. Some historical corpora, including certain cartularies or collections of texts produced within a relatively coherent institutional or linguistic context, may also display these characteristics, although this depends strongly on their internal heterogeneity and on how the texts are segmented. In Table 5, we present an excerpt from the topics generated through LDA on Henry's chronicle.

Table 5. Examples of topics identified by LDA (most significant words, number of topics = 10).

1	mare	ecclesia	mater	magnus	dominus	christus	episcopus
2	exercitus	paganus	peregrinus	fides	frater	lyvones	episcopus
3	seculum	fides	verbum	abbas	legatus	christus	rego
4	expeditio	exercitus	villa	terra	equus	vir	castrum

Once the topics are generated, they can be used as new variables for the study of vocabulary. This drastically reduces the size of the variable space from more than 10,000 word forms in this case to only ten topics, thus making visualisations possible, albeit with a significant loss of information. We can, for example, study the intensity of topics over time or analyse the correlations between them.

Let us now consider the distribution of our first four topics across Henry's chronicle (Figure 3). In contrast to the topics produced through clustering, we observe much less noise: a small number of

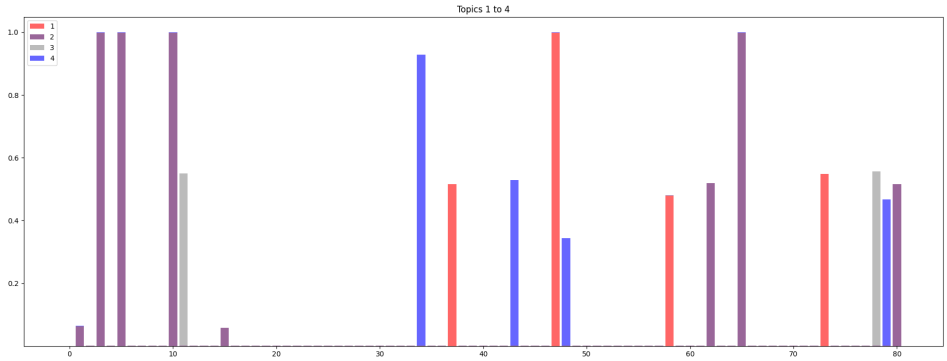


Figure 3. Comparative frequency of some LDA topics across consecutive two-page segments.

topics predominate within each textual unit, and their distribution is more clearly localised. Topic 2, for instance, which contains primarily vocabulary relating to crusading armies, is especially prominent during the initial expedition (HCL II–X) and during the conquest of Estonia (HCL XXIII and XXIX). By contrast, topic 1 appears mainly in sermons in the second part of the chronicle, while topic 3 concerns two specific episodes in which negotiators are sent between rival powers (HCL IX and XXIX).

EMBEDDINGS

One way of moving beyond direct lexical comparison is to represent words according to the contexts in which they occur. Embedding methods provide such representations, although they do not in themselves eliminate differences of language, genre, authorship, or style.

An embedding is a numerical representation of a linguistic unit (such as a word, sub-word, sentence, or passage) as a vector in a multidimensional space. Different model families construct such representations in different ways. Static word-embedding models, such as Word2Vec and GloVe, assign each word type a single vector learned from patterns of co-occurrence in a corpus. Contextual language models, including transformer-based models, instead generate representations that depend on the surrounding text, so that the same word may receive different vectors in different contexts. Encoder–decoder models constitute a further architecture in which such internal representations are used to transform one sequence into another, for example in machine translation.

Instead of treating each word[16] as an isolated symbol, embeddings place words in a continuous vector space where their positions reflect semantic and syntactic relationships. In the following examples we use a static word-embedding approach. Word2Vec learns vector representations by exploiting patterns of word co-occurrence in a corpus, so that words occurring in similar linguistic contexts tend to occupy nearby regions of the resulting vector space.[17] In Figures 4 and 5, we show examples of such embeddings applied to the vocabulary of Henry's chronicle and the *Livonian Rhymed Chronicle*, respectively.

A latent space is the abstract, multidimensional space in which these vectors reside. The term 'latent' refers to the fact that the dimensions of this space are not directly interpretable; they are learned features that capture underlying patterns in the data rather than explicit, human-defined properties. Each dimension may encode a mixture of linguistic traits, such as formality, plurality, or semantic category, although not in a way that can be clearly labelled. What matters is not the meaning of any single dimension, but the geometric relationships between points. Distance and direction in this space correspond to similarity and relational structure.

t-SNE (t-distributed stochastic neighbour embedding) is a widely used method for visualising high-dimensional data in two dimensions. Its principal strength lies in preserving local neighbourhood relationships: points that are close in the original high-dimensional space will often remain close in the two-dimensional projection. Unlike PCA (principal component analysis), t-SNE is non-linear and is therefore capable of representing complex local structures that may be difficult to visualise using linear projection. This advantage comes with an important limitation, however: distances between widely separated points, the relative positions of clusters, and even the apparent size of clusters in a t-SNE plot should not be interpreted as faithful representations of the original space. The visualisations in Figures 4 and 5 should therefore be

16 Whereas words can be directly embedded in a latent space, modern language models generally operate on tokens rather than words. Tokenisation differs fundamentally from lemmatisation: whereas lemmatisation reduces each word to a common lexical form while leaving the syntactic structure of the sentence unchanged, tokenisation represents a sentence as a sequence of tokens. A token is an abstract unit that typically corresponds to part of a word, although it may sometimes correspond to an entire word or, in some cases, to several words. Crucially, the division of a word into one, two, or even several tokens is not guided by grammatical or linguistic considerations, but is determined primarily by the statistical distribution of linguistic units in the corpus on which the model is trained.

17 L. van der Maaten, G. Hinton. Visualizing Data Using t-SNE. – Journal of Machine Learning Research, 2008, 9, 2579–2605. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf; T. Mikolov, K. Chen, G. S. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. – International Conference on Learning Representations, 2013. https://doi.org/10.48550/arXiv.1301.3781.

read primarily as exploratory representations of local proximity rather than as maps of the global structure of the embedding space.

During training, the model processes text and attempts to predict relationships between words. Each word begins as a random vector, but through repeated exposure to text and iterative optimisation, the

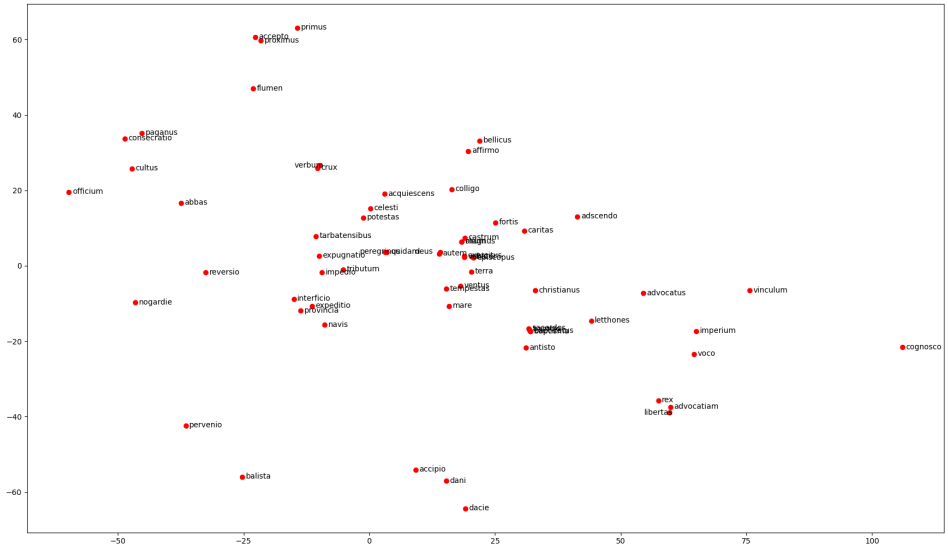


Figure 4. t-SNE projection of selected words from HCL. Local proximity may indicate similarity in distributional context; global distances in the two-dimensional projection should not be interpreted directly.

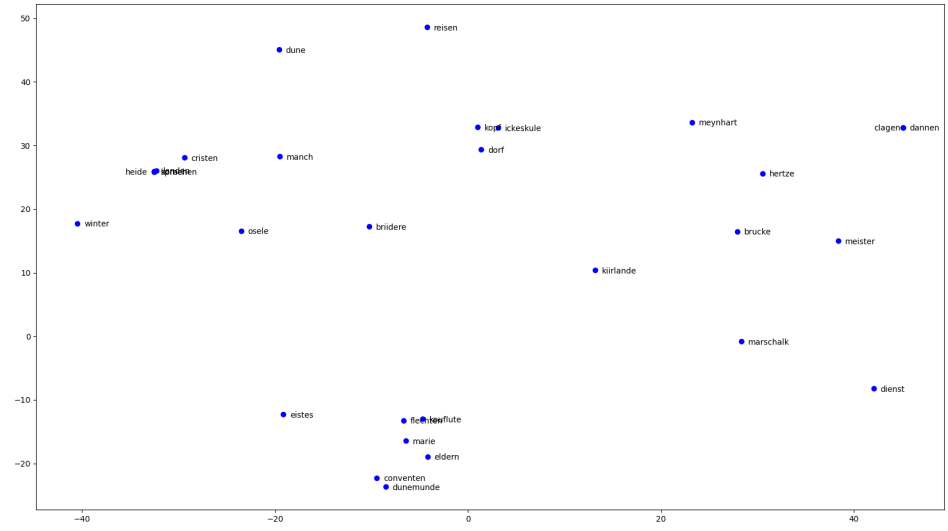


Figure 5. t-SNE projection of selected words from RC.

vectors gradually adjust to capture meaningful patterns. Over time, coherent structures emerge: words such as *conventen* and *dunemunde* end up close to each other in the space, and relationships such as gender or tense distinctions can be encoded as consistent vector offsets. If we had sufficient material for more extensive training and time to explore the results in greater depth, we might even hope to discover arithmetic relationships similar to those observed in English word embeddings, where subtracting the vector for 'man' from that for 'king' and adding the vector for 'woman' yields a vector close to that for 'queen'. In a similar spirit, one could imagine relations such as *rex* + *ecclesia* − *miles* ~ *episcopus*, although this does not currently emerge in our data.

It is indeed important to keep in mind that, while it is possible to train embeddings on a small corpus for the purpose of clustering, training a model capable of more complex tasks such as question answering requires significantly more resources and training data. Pretrained language models capable of processing Latin are increasingly available, although their performance varies according to period, genre and task. Comparable resources for Middle High German remain more limited, particularly for the kinds of context-sensitive tasks considered here.

SIMILARITIES BETWEEN CHRONICLES

The problem becomes more difficult when we attempt to compare passages across the two chronicles. A direct lexicometric comparison is strongly affected by language and vocabulary. Even when both chronicles are analysed through their English translations,[18] TF-IDF tends largely to separate the two texts rather than identify passages describing comparable events (Figure 6). The following experiment therefore adopts a different strategy. Rather than directly comparing independently trained Latin and Middle High German embeddings, we construct thematic representations separately for the two chronicles and then identify approximate correspondences between topics on the basis of their most significant words. The resulting comparison is thus partly model-based and partly researcher-mediated: the models identify patterns within each corpus, while the correspondence between thematic

18 The Chronicle of Henry of Livonia. Trans. by J. A. Brundage; The Livonian Rhymed Chronicle. Trans. by J. C. Smith, W. L. Urban. Indiana University, Bloomington, 1977.

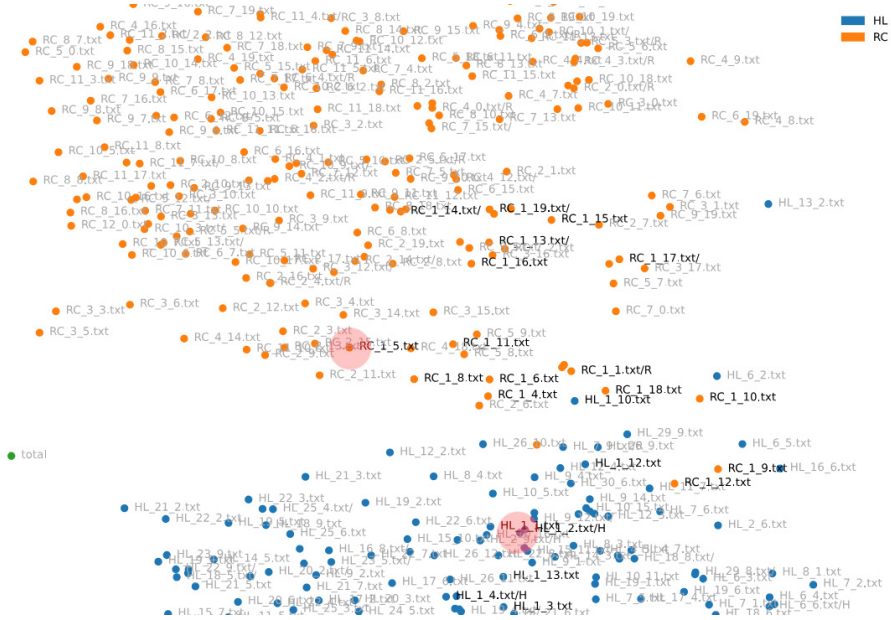


Figure 6. t-SNE projection of chapters from Henry's chronicle and the Livonian Rhymed Chronicle based on word-level TF-IDF in the English translations. The pink circles show the position of two texts describing the same event (the arrival of Meinhard).

structures across the two languages is established through historical and linguistic interpretation.

Once these approximate thematic correspondences have been established, passages from the two chronicles can be compared in terms of their topic composition. Most passages remain associated primarily with one chronicle, but several cases display comparable thematic profiles across the two texts. These similarities should not be understood as having been discovered through a language-neutral embedding space; rather, they emerge from a combination of statistical modelling within each corpus and scholarly alignment between the resulting topics.

The passages represented by neighbouring red and blue points in Figure 7 deal with warfare against Russian princes, as illustrated by HCL XXII.2 (pp. 148–149).

If we look closely at the topic distribution in this chapter, we see that topic 4 accounts for more than 90 per cent of the estimated topic mixture in this passage, which deals with warfare, the description of armies, and Russian powers. Similar topics are omnipresent in the *Rhymed Chronicle*, hence the association.

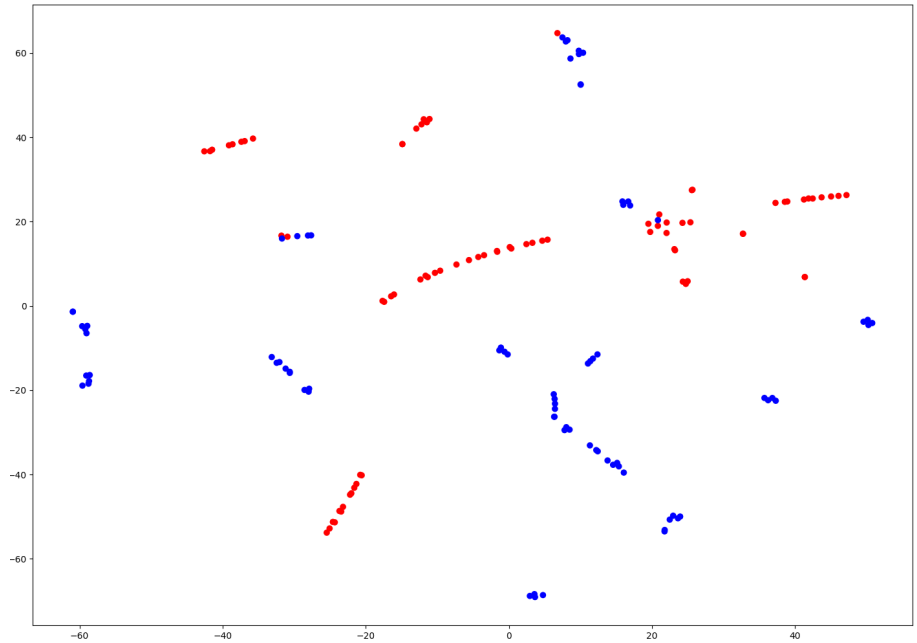


Figure 7. t-SNE projection of pages from Henry's chronicle (red) and the Livonian Rhymed Chronicle (blue) based on thematically aligned LDA topic representations. Topic models were generated separately for the two corpora; cross-language correspondences between topics were established manually.

VERIFICATION OF RESULTS

Evaluating the results produced by a statistical tool is a problem as old as statistics itself. An important part of the data scientist's work consists in designing and applying control procedures throughout the analytical process, both to assess and improve the results and, ultimately, to provide an estimate of the validity of the final outcome. This process rests essentially on two pillars: metrics and control samples.

Data-science textbooks provide well-established principles for designing robust evaluation metrics. A simple accuracy score, for instance, can be misleading when the classes are unbalanced. A rate of 95 per cent correct classifications may not be particularly impressive when distinguishing sentences that do or do not refer to King Philip of Swabia in Henry's chronicle (just answer 'no' to all and you have it), whereas it would be a perfectly respectable result for identifying paragraphs containing a reference to Holy Scripture.

It is often easier to identify false positives (is the selected paragraph actually about baptism in Vironia?) than false negatives (can I be sure

that no passage dealing with siegecraft has been discarded?). The former requires only a quick examination of the selected passage, whereas the latter may require an exhaustive inspection of the material that was not selected. Since both false positives and false negatives need to be assessed with equal care, the construction of control samples is essential. Such samples must be large enough to be representative, yet small enough that a reliable ground truth can be established with a reasonable investment of effort.

In data analysis, many evaluation metrics are automated for the sake of efficiency. In historical research, however, scholars will often rely on what might be called 'manual' metrics, for instance, their own assessment of the quality of a translation.[19] Although manual evaluation is more time-consuming, it does not fundamentally alter the principles governing the construction of samples and metrics.

Once an evaluation framework has been established, the algorithm can be modified and re-run iteratively until its performance on the control sample reaches a satisfactory level. If substantial changes are made to the model or its design during this process, however, it is safer to revise the control sample periodically, in order to reduce the risk of overfitting to it. Finally, a separate test sample, which has not been used at any stage of the development process, can be used to compute the final evaluation metrics and provide an independent assessment of the results.

STATISTICAL MODELS AS PARATEXTUAL TOOLS

Throughout the modern period, archivists and scholars have sought to develop tools such as printed editions of manuscripts, compilations, marginal notes, indexes and regesta. All were designed to spare future researchers the need to begin the study of manuscripts from scratch and to facilitate the gradual construction of a collective understanding of historical sources.

Emerging progressively from medieval glosses and early modern scholarly conventions, these tools became increasingly standardised between the seventeenth and nineteenth centuries, enabling historians to organise vast amounts of information, reference sources, and construct

19 More complex architectures can blur this distinction, particularly when language models are used to evaluate the outputs of other language models. Such approaches may raise difficult questions even for machine-learning specialists, and most historians are unlikely to need to venture into this particularly treacherous territory.

arguments with greater transparency and rigour. Indexes allowed readers to access themes non-linearly, while footnotes and references created a traceable network of verifiable evidence, provided that the original materials were preserved. In many ways, these paratextual systems externalised the historian's method, making visible the processes of selection, citation, and interpretation.

This has helped historians work with increasingly large corpora while grounding their claims in systematic surveys of the sources rather than impressions derived from reading a limited number of texts. However, it was not without drawbacks. Researchers have long emphasised the importance of maintaining direct contact with original materials: studying palaeography in order to read manuscripts directly, treating regesta as second-hand material, systematically verifying comments and references, and publishing regularly updated editions intended to correct errors and incorporate advances in scholarly knowledge. Nevertheless, very few historians today would argue for abandoning the tools of critical editions altogether.

The large-scale digitisation of textual materials and the rise of the internet can be understood as a contemporary parallel to the tradition of critical editions of sources, extending their core aims while transforming their scale and accessibility. Digital archives further the purpose of earlier critical editions by providing searchable transcriptions and layered metadata that allow users to trace provenance, compare versions, and examine editorial decisions. Online platforms also make it possible to juxtapose multiple editions or manuscripts instantly. As with earlier paratextual tools, the promise of accessibility and critical transparency depends on maintaining rigorous standards, as well as on developing digital literacy within the scholarly community, in order to avoid drawing unfounded conclusions from partially understood materials.

In this context, statistical methods – and especially more recent approaches based on large language models – occupy a particular position, since it is virtually impossible for the vast majority of historians to acquire the highly specialised skills required to fully understand and critically assess the way their outputs are produced. While palaeography, Latin, archival classification systems, XML encoding, and even Python scripting can be integrated into a historian's initial training to the point of making the use of digital materials both safe and fruitful, statistical methods can only be mastered up to a certain point. Even specialists working in the field of machine learning must constantly update their knowledge, navigating a landscape in which the dissemination of research is often

hindered by economic competition and by the difficulty of maintaining an accessible and up-to-date body of knowledge within the community.

However, indexes and marginal notes were certainly very useful to historians who could not access original manuscripts directly. As long as they were familiar with primary sources in general and understood the process of critical editing, there was little need for them to decipher the particularly difficult handwriting of a given fifteenth-century lawmaker when working from a printed or digital transcription. In a similar way, we argue that a solid understanding of basic statistical tools may be sufficient for scholars to make effective use of enhanced textual editions that incorporate more advanced statistical methods, provided these tools are designed so that the additional information remains easy to interpret and verify.

ML-ENRICHED METADATA

Many readers are probably already familiar, if not with the specific conventions of TEI-XML, then at least more generally with the use of tags to integrate metadata into the digital edition of a historical source. This practice is a direct evolution of the footnotes and marginal comments of earlier printed editions, designed to be revealed or hidden on demand, and to be easily searchable using software tools. These tags may be used to indicate information about the original manuscript's appearance, its current location, or its successive versions. They can also provide references to secondary materials or help categorise texts within a broader corpus, among many other uses.

```
< chapter type="chapter" n="1.4" >
< topics allocation="lda8tfidf42" >
  < topic id="1" p="0.7"/ >
  < topic id="5" p="0.3"/ >
< /topics >
< head > (4) < /head >
< p > And in the same village < fn > Ükskülll, nowadays Ikšķile < /fn > Ylo, the father of Kulewene,
and Viezo, the father of Alo, were the first to be baptized, while the others followed in their turn. < p >
...
< allocation id="lda8tfidf42" method="LDA" n="8" >
  < topic id="1" >
    < word text="baptized" p="0.632"/ >
    < word text="holy" p="0.611"/ >
    < word text="baptism" p="0.598"/ >
  < /topic >
  < !-- Additional topics omitted -- >
< /allocation >
```

Figure 8. Example of a metadata-enriched edition using custom XML markup to link a chapter to topic-modelling results.

These metadata-enriched editions may represent the most effective integration of statistical tools into a historian's engagement with sources. Consider, for instance, a scenario in which a data scientist, in collaboration with a historian specialising in the relevant field, produces a topic model of a given corpus. The results of this modelling can then be integrated into a digital edition in the form of metadata (Figure 8).

In this example, we use the tag "topics" to indicate the existence of such an analysis, with the attribute "allocation" referring to a methodological section where the details are clearly specified. In this section, the method and its parameters are documented in order to ensure the process is reproducible. In addition, the full results are presented within the same methodological framework: lists of words for each topic, along with the model's confidence in the assignment of each item.

Thus, a scholar interested in identifying all texts in a corpus dealing with, for instance, medieval piracy or the ordination of canons could use these topic-based metadata through a search engine in much the same way as they would use a traditional index. Conversely, another scholar with deeper mathematical literacy and a specific interest in a topic not captured by the initial metadata could adjust the parameters until a topic relevant to their research emerges from the process.

COLLABORATIVE GENERATION OF METADATA WITH LLMS

When it comes to generating metadata for historians, topics – whether produced through frequency-based clustering or Latent Dirichlet Allocation – represent only one form of computationally generated metadata. For most uses, they may already be outdated or superseded by more recent methods. Beyond the validity of their theoretical framework, their main limitation is that their output is difficult to adapt to the specific questions a researcher may wish to ask, much like traditional indexes, which were often restricted to personal names, place names, or core concepts. By contrast, large language models (LLMs) offer historians a powerful means of generating and organising metadata at scale.

Large language models can automatically extract key attributes such as dates, locations, personal names, topics, and genres, even when the original documents are unstructured or written in archaic languages. Their ability to perform such tasks derives from large-scale language modelling, through which they learn contextual patterns and

associations that can be mobilised through natural-language instructions. In addition, the use of natural language allows researchers to be far more directive in shaping the form and scope of the output. For instance, it is relatively straightforward to ask a model to produce a table of all sieges mentioned in a chronicle, with precise references to the relevant passages. The accuracy and verifiability of the resulting tables, however, remain separate questions. As a simple illustration, we asked a locally run Llama 3 (8B) model to extract keywords, a task at which such models perform reasonably well, since it primarily requires an understanding of local context, and to append them as tags at the end of paragraphs in Henry's chronicle:

> Anno domini 1198 venerabilis Albertus, Bremensis canonicus, in episcopum consecratur. Post consecrationem estate proxima Gothlandiam vadit et ibidem circa quingentos viros signo crucis ad eundum in Lyvoniam signat. Inde per Daciam transiens munera regis Canuti et ducis Woldemari et Absolonis archiepiscopi recepit.
>
> <keywords>1198, consecration of bishop, Albert of Bremen, crusading recruitment, Gotland, Livonia mission, Denmark, King Canute, Duke Valdemar, Archbishop Absalon, medieval church networks</keywords>
>
> Reversus in Theuthoniam in natali domini Magdeburch plures signat. Ubi rex Philippus cum uxore coronatur. Et coram eodem rege in sentencia queritur, si bona in Lyvoniam peregrinancium sub tuicione pape ponantur, sicut eorum, qui Ierosolimam vadunt. Responsum vero est ea sub protectione apostolici comprehendi, qui peregrinacionem Lyvonie in plenariam peccaminum remissionem iniungens, vie coequavit Ierosolimitane.
>
> <keywords> 1199, church assembly, recruitment of crusaders, coronation of king, papal protection, legal status of pilgrims, Livonian crusade, comparison to Jerusalem pilgrimage, indulgences, ecclesiastical authority </keywords>
>
> Anno secundo sui episcopatus cum comite Conrado de Tremonia et Harberto de Yborch cum multis peregrinis in Lyvoniam vadit, habens secum in comitatu 23 naves.
>
> <keywords> 1200, second episcopal year, expedition to Livonia, Count Conrad of Dortmund, Harbert of Iburch, crusading fleet, maritime campaign, medieval colonisation, missionary expansion </keywords>

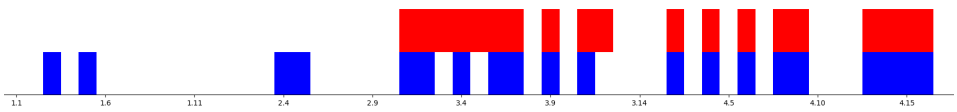


Figure 9. Chapters in Henry's chronicle assigned the keyword 'seafaring' in two runs of an LLM using slightly different prompts.

Although they perform reasonably well overall, models may introduce inaccuracies, such as confusion over dates, or overly confident

interpretations, for instance, assigning "colonisation" as a keyword in the last paragraph, which could be debated but is certainly not supported by that specific passage. More subtly, they can produce outputs that appear broadly appropriate but contain irrelevant or weakly informative items; for example, "second episcopal year" would probably not constitute a meaningful keyword. Figure 9 illustrates how the chapters assigned a particular keyword can vary with the prompt.

While topic models, for instance, provide probabilities of classification that can be useful for identifying errors, raw token probabilities generated by large language models do not provide straightforward or well-calibrated confidence estimates for the factual correctness of generated metadata. Moreover, proprietary systems may not expose these probabilities to users at all. As a result, such probabilities cannot by themselves be used as reliable indicators of whether generated metadata is historically accurate, and the outputs therefore require independent validation. Conversely, these models can be highly effective when used to generate metadata within a well-defined and controlled process.

While a single historian or data scientist could easily be misled by an LLM's output, it is important to recognise that the construction of trustworthy metadata is a collective and incremental process. The *Liv-, Esth- und Curländisches Urkundenbuch*, initiated by Friedrich Georg von Bunge, was corrected, curated, and expanded by many scholars over the following century. Leonid Arbusow and Friedrich Benninghoven, in particular, systematically revised its coverage of the period up to 1300,[20] but many additional pieces of information have since been disseminated through the scholarly literature. Before the rise of language models, digitisation and collaborative editions already helped accelerate this process, enabling proofreading, the inclusion of new materials, and the addition of references to occur much more rapidly under the supervision of the scholarly community, although the mechanisms that make such supervision reliable remain, in themselves, the subject of extensive debate.

Large language models amplify both the possibilities of these tools and their associated challenges. The time required for generation – even including prompt design and preliminary review of small samples – is generally much shorter than the time required for validation. Depending on the specificity of the language, the size of the corpus, the need to

20 Liv-, est- und kurländische Urkundenregesten bis zum Jahre 1300. Hrsg. von F. G. v. Bunge, mit Ergänzungen von L. Arbusow jun., neu herausgegeben von F. Benninghoven, Geleitwort von P. Johansen. Hamburg, 1959.

incorporate external information or the possibility of framing the question in precise terms, the use of LLMs for metadata generation may or may not be justified.

Moreover, language models, as well as more conventional software and statistical tools, can be useful in the incremental improvement of such metadata. For the specific task of generating keywords alone, a non-exhaustive list of applications includes standardising keyword sets, clustering similar names, removing hapaxes, and compiling references. The metadata generated can thus be processed using a range of established digital analysis tools.

At a later stage in the workflow, scholars can use an LLM with a large data repository more effectively if the underlying texts have been enriched with metadata. In 2024, retrieval-augmented generation (RAG) was the cornerstone of such techniques, although many architectures combining the model itself with function calling and various search engines can now be used, particularly with models capable of handling large input contexts. In any case, these systems generally benefit from curated metadata, much as classical search tools do. Thus, alongside traditional software, LLMs can be useful at both ends of the process, provided that strong mechanisms of control are maintained in each case.

Before concluding, it is worth emphasising that this article has explored only the tip of the iceberg when it comes to LLM-assisted research in history, which is a very rapidly expanding topic. A wide range of applications are currently being developed and critically examined, many of which the author has not tried himself at the time of writing. Readers interested in exploring this rapidly evolving field may wish to begin with Andres Karjus's review of some of the most recent and promising experiments in the area.[21]

CONCLUSION

This article has examined what different levels of computational text analysis can contribute to the study of medieval historical texts, while also emphasising the methodological limitations that accompany them. Applied to the *Chronicle of Henry of Livonia* and the *Livonian Rhymed Chronicle*, the methods discussed here show that computational analysis

21 A. Karjus. Machine-assisted quantitizing designs: augmenting humanities and social sciences with artificial intelligence. – Humanities and Social Sciences Communications, 2025, 12, 277. https://doi.org/10.1057/s41599-025-04503-w.

can identify recurring patterns, reveal relationships between passages, support carefully designed comparisons between texts written in different languages, and assist in the production of searchable metadata. At the same time, the experiments demonstrate that the evidential value of these results depends heavily on corpus size and quality, textual segmentation, linguistic variation, model design, and the procedures used to validate the output.

These dependencies become more important as the methods themselves grow more complex. Relatively transparent techniques such as frequency analysis and clustering allow researchers to inspect fairly directly how results are produced, even if they remain sensitive to lexical ambiguity and arbitrary textual divisions. Topic models, embeddings, and transformer-based language models can capture more complex patterns, but they also introduce additional layers of modelling and interpretation. Their outputs therefore require correspondingly stronger forms of methodological documentation, validation, and scholarly mediation. Increasing computational sophistication should not be confused with increasing evidential autonomy.

The emergence of transformers and large language models further expands these possibilities by enabling, among other applications, the automatic extraction of metadata, semantic tagging, and complex textual queries. Nevertheless, these systems also introduce new risks. Their outputs may appear coherent and authoritative while containing subtle inaccuracies, anachronisms, or unsupported interpretations. Because the internal logic of such models is often opaque even to specialists, the historian's task cannot simply become one of accepting algorithmic results. Instead, we must approach these tools critically, understanding that their value lies not in producing definitive interpretations but in assisting the construction, organisation, and exploration of historical corpora.

For this reason, we argue that the most productive role of machine learning in historical research is not as a substitute for scholarship, but as part of a collaborative and transparent paratextual infrastructure. Just as indexes, critical editions, footnotes, and regesta transformed historical practice in earlier centuries, ML-generated metadata and digitally enriched editions can facilitate the collective accumulation, organisation, and revision of knowledge today. Their usefulness depends on rigorous documentation, reproducibility, clear provenance, and continuous scholarly verification. Statistical methods are most useful

when integrated into workflows in which historians, linguists, and data scientists jointly design, evaluate, and refine the tools being used.

Ultimately, the responsible integration of computational methods into medieval studies depends on balancing technological possibilities with methodological control. Historians do not need to become machine-learning specialists, but they do need sufficient statistical literacy to understand the assumptions and limitations of the systems they use. Likewise, developers of computational tools must prioritise interpretability, accessibility, and scholarly transparency over technical sophistication alone. Used critically and collaboratively, such methods can greatly enrich the study of medieval texts, while contributing to more accessible, interconnected, and verifiable forms of historical knowledge.

Acknowledgements

The author used ChatGPT to proofread the English throughout the text and to generate initial drafts of the abstract and summary. In the section 'Collaborative Generation of Metadata with LLMs', he used a local copy of Llama 3 (8B). All other analyses and data processing were carried out using the scikit-learn library.

AJALOOLANE JA TÄNAPÄEVASED STATISTILISED MEETODID: RAKENDUSI 13. SAJANDI LIIVIMAA KROONIKATE UURIMISEL

Nicolas Bourgeois

Artiklis uuritakse, kuidas saab tänapäevaseid statistilisi keeletöötlusmeetodeid rakendada keskaegsete ajalootekstide uurimisel, keskendudes eeskätt 13. sajandi Liivimaa kahele peamisele kroonikale: ladinakeelsele Henriku Liivimaa kroonikale ja keskkõrgsaksakeelsele Liivimaa vanemale riimkroonikale. Autori eesmärk ei ole piirduda tehniliste meetodite tutvustamisega, vaid arutleda ka praktiliste ja tunnetuslike probleemide üle, millega ajaloolased tehisintellekti tööriistu kasutades kokku puutuvad.

Sissejuhatuses osutab autor, et tehisintellekti, eriti loomuliku keele töötluse (NLP), kiire areng on tekitanud märgatava lõhe selle vahel, mida need süsteemid suudavad teha, ja selle vahel, kuivõrd teadlased

nende süsteemide toimimisest tegelikult aru saavad. Ajaloolased võivad neid tööriistu keerukuse tõttu kas umbusaldada või, vastupidi, usaldada liigselt tulemusi, mille kujunemist ja usaldusväärsust nad piisavalt hinnata ei oska. Seetõttu peab autor vajalikuks kasutada eeskätt selliseid tööriistu, mille toimimine on võimalikult läbipaistev ja tulemused tõlgendatavad ning mida saab lõimida koostöisele uurimistööle tuginevatesse teaduspraktikatesse.

Esimeses, metodoloogilises osas käsitletakse leksikomeetriat. Lihtne sõnasageduste analüüs võib olla eksploratiivses uurimistöös kasulik, kuid selle võimalused on piiratud. Sama sõna eri vorme võidakse käsitada eri sõnadena, samal ajal kui üks ja sama sõna võib eri kontekstides kanda erinevat tähendust. Autor näitab, kuidas sageli koos esinevate sõnade klasterdamise abil on võimalik tuvastada näiteks sõjapidamise, religiooni või meresõiduga seotud teemad. Sellised meetodid võivad aga anda ka juhuslikke või eksitavaid tulemusi, sest sõnad paigutatakse jäigalt ühte kategooriasse ning teksti meelevaldne jaotamine väiksemateks osadeks võib moonutada sõnadevahelisi semantilisi seoseid.

Seejärel tutvustatakse artiklis latentse Dirichlet' jaotuse meetodit (*Latent Dirichlet Allocation*, LDA), mida kasutatakse tõenäosuslikus teemamudeldamises. Erinevalt lihtsast klasterdamisest lähtub LDA eeldusest, et iga tekst sisaldab korraga mitut varjatud teemat ning iga teemat iseloomustab omakorda sõnade tõenäosusjaotus. Nii võib üks ja sama sõna kuuluda sõltuvalt kontekstist erinevatesse semantilistesse väljadesse. Henriku kroonikale rakendatuna võimaldab LDA eristada sidusamaid ja tekstis paremini lokaliseeritavaid teemasid ning jälgida seeläbi näiteks ristisõdade, diplomaatia või jutlustamisega seotud ainest kogu kroonika ulatuses. Samas toimib see meetod kõige paremini suurte ja suhteliselt ühetaoliste tekstikorpuste puhul, mida keskaja-uuringutes sageli kasutada ei ole.

Järgmises osas käsitletakse sõnavektoreid ja latentseid ruume. Sellised mudelid nagu Word2Vec esitavad sõnu vektoritena mitmemõõtmelises ruumis, kus tähenduslik sarnasus väljendub sõnade geomeetrilises läheduses. Näiteks võimaldab t-SNE neid seoseid ka visuaalselt esitada. Eri keeltes tekstide võrdlemisel eeldab sõnavektorite kasutamine eri keelte vektorruumide joondamist või ühist mitmekeelset mudelit. Artiklis esitatud kahe kroonika võrdlus põhineb eraldi loodud teemamudelite tulemuste käsitsi joondamisel. Nii tuvastatakse sarnase teemakoostisega tekstikohti, näiteks sõdu vene vürstide vastu käsitlevaid lõike.

Artiklis eristatakse tokeniseerimist ja lemmatiseerimist. Tokeniseerimine jagab teksti mudeli töödeldavateks üksusteks, näiteks sõnadeks, sõnaosadeks või märkideks; lemmatiseerimine koondab sama sõna eri vormid ühise algvormi alla. Keskaegsete keelte puhul võivad nii kirjaviis kui ka sõnavormid suuresti varieeruda, mis võib raskendada mõlemat töötlusetappi ning mõjutada analüüsi tulemusi.

Artikli lõpuosas vaadeldakse tehisintellekti ja statistilisi mudeleid eeskätt paratekstiliste tööriistadena, mida võib võrrelda registrite, glosside ja kriitiliste väljaannetega, millele ajaloolased on oma töös traditsiooniliselt toetunud. Autori hinnangul ei pea ajaloolane tehisintellektisüsteemidest kasusaamiseks täielikult valdama nende matemaatilist ülesehitust, nii nagu trükitud allikaväljaande kasutamine ei eelda tingimata paleograafia täielikku valdamist. Praktiliseks vahevõimaluseks peetakse seetõttu hoolikalt koostatud digitaalseid väljaandeid, mida on rikastatud läbipaistvalt esitatud metaandmetega.

Artiklis tehakse ettepanek lõimida masinõppe abil loodud metaandmed teaduslikesse väljaannetesse läbipaistvate märgendussüsteemide abil. Teemamudelite ja suurte keelemudelite (LLM) abil saab automaatselt luua märksõnu, liigitada tekstilõike ning tuvastada selliseid olemeid nagu isikud, kohad ja sündmused. Autor rõhutab siiski, et tulemusi ei saa käsitleda lõplikena ja neid tuleb alati teaduslikult kontrollida. Suured keelemudelid võivad teha raskesti märgatavaid vigu, pakkuda liiga kaugele minevaid tõlgendusi või anda eksitavaid liigitusi. Selliseid eksimusi võib olla keeruline avastada, sest mudelite sisemine toimimine ei ole kasutajale läbipaistev.

Kokkuvõttes rõhutab artikkel, et tehisintellekti tööriistad ei peaks asendama kriitilist ajaloouurimist, vaid toetama ajaloolise materjali indekseerimist, annoteerimist ja korrastamist kui koostöist teaduspraktikat. Tehisintellekti tööriistade suurim väärtus ei seisne autoriteetsete ja iseseisvate tõlgenduste loomises, vaid kollektiivsete, läbipaistvate ja pidevalt täiendatavate teadustaristute kujundamise toetamises.