Understanding the System in its Entirety?

Digital Corpora, Language Models and Low Resource Languages

Gabriel Viehhauser

1. Texts from the pre-printing era as epistemic objects in the Digital Humanities

Since the institutionalization of the discipline at the beginning of the millennium, it has become common sense that one of the core tasks of Digital Humanities is not only to use computer-assisted methods to answer research questions from the humanities but also to reflect on their application. In fact, digital philology and digital literary studies (a traditional core area of the Digital Humanities) in particular have produced no shortage of work reflecting on how digital methods can be used meaningfully to do justice to literary objects despite (or precisely because of) the formalization that is inevitable in the computer field.[1] However, it seems that the observation that digitization and methodology also affect the constitution of the objects has been given comparatively fewer consideration.[2]

The fact that texts are texts not in a ›self-evident‹ way but rather as the results of fundamental conceptualizations of textuality may be more relevant to those fields of literary studies whose objects of investigation are not (or no longer) subject to the paradigms of the print age and modern book production, thus to medieval literary studies (and also more and more in research that focusses on born digital literature). In fact, questions such as how textuality is constituted outside the book medium, how the concept of literature (beyond a differentiated literary industry)

1 Cf. e.g. Reiter, Nils/Pichler, Axel/Kuhn, Jonas (2020): Reflektierte algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt, Berlin/Boston: De Gruyter; the »Debates in the Digital Humanities« series (Online: https://dhdebates.gc.cuny.edu, last access: 21.09.2025); or the debate on Da's critique in Da, Nan Z. (2019): »The Computational Case against Computational Literary Studies«, in: Critical Inquiry 45, pp. 601–639.

2 Cf. Limpinsel, Mirco (2016): »Was bedeutet die Digitalisierung für den Gegenstand der Literaturwissenschaft?«, in: Zeitschrift für digitale Geisteswissenschaften. Online: https://zfdg.de/2016_009 (last access: 21.09.2025).

should be conceived, and how far it stretches (›narrow‹ vs. ›broad concept of literature‹) have been extensively discussed in medieval philology long before the advent of digital technology, because they impose themselves due to the media difference between the manuscript and the print paradigm—even though medieval characteristics such as orality and performance were also overlooked for a long time, mainly because medieval texts were conceived as books, especially in print editions.[3]

2. Varieties of the virtualization of medieval literature

2.1 Modelling the material text: Facsimiles

However, as has been shown by Lechtermann and Stock, the digital transformation adds a new layer of disruptions in the construction of ›literature‹ and ›text‹ as objects of medieval literary studies.[4] In the case of pre-modern texts, Lechtermann and Stock observe two contradictory effects of digitalization:[5] In its first variety, digitization emphasises the material dimension of the text. Digital facsimiles of manuscripts, for example, enable the examination of the materiality of medieval texts and the conduct of meticulous palaeographic and codicological studies of subtle characteristics of handwriting. Compared to the mere lettering of a print edition, they thus open up a view of the material embedding of the text and, to a certain extent, reverse the transformation of the manuscript text into book form brought about by the print paradigm.[6] But because these facsimiles are always only images and virtual conceptualizations of the codex, not the object itself, they can never comprehensively catch up with the materiality of the manuscript. A digital

3 Cf. e.g. Baisch, Martin (2013): »Textualität – Materialität – Materialität – Textualität. Zugänge zum mittelalterlichen Text.«, in: Literaturwissenschaftliches Jahrbuch 54, pp. 9–30; Strohschneider, Peter (1997): »Situationen des Textes: Okkasionelle Bemerkungen zur ›New Philology‹«, in: Helmut Tervooren/Horst Wenzel (eds.): Philologie als Textwissenschaft. Alte und neue Horizonte, Berlin: Schmidt, pp. 62–86; Strohschneider, Peter (1999): »Textualität der mittelalterlichen Literatur: Eine Problemskizze am Beispiel des ›Wartburgkrieges‹«, in: Jan-Dirk Müller/Horst Wenzel (eds.): Mittelalter: Neue Wege durch einen alten Kontinent, Stuttgart/Leipzig: Hirzel, p. 19–41; Strohschneider, Peter (2014): Höfische Textgeschichten. Über Selbstentwürfe vormoderner Literatur, Heidelberg: Winter.

4 Lechtermann, Christina/Stock, Markus (2020): »Virtuelle Philologie«, in: Dawid Kasprowicz/Stefan Rieger (eds.): Handbuch Virtualität. Wiesbaden: Springer, pp. 425–454; Lechtermann, Christina/Stock, Markus (2021): »Virtuelle Textkonstitutionen: Die Philologie und ihre mittelalterlichen Objekte«, in: Stefan Rieger/Armin Schäfer/Anna Tuschling (eds.): Virtuelle Lebenswelten: Körper – Räume – Affekte, Berlin/Boston: De Gruyter, pp. 63–86.

5 Lechtermann/Stock: Virtuelle Textkonstitutionen, p. 71.

6 For the moment, I will leave aside the fact that manuscripts themselves can in many cases also be second-level representations of texts that were originally intended for oral presentation.

facsimile misses the haptic and other sensory qualities of the codex as an object, and the screen display levels the latter's three-dimensionality.[7]

Thus, facsimiles could be considered models of the text, and it is in the nature of models to highlight certain aspects of reality while reducing others. This reduction is necessary in order to create knowledge and to not get lost in the flood of nuances. Models therefore always serve a specific purpose.[8]

Accordingly, it is precisely the possibility of manipulating the facsimile that separates the virtual from the ›real‹ material object. One example of this would be the ability to change the size of facsimiles as desired, which offers a flexible display for analysis but may also obscure the true dimensions of the manuscript object. In digital repositories, manuscript pages are usually levelled regardless of the actual size of the codices in order to optimize the screen display, so that the sense of different proportions is lost in virtualization.[9] Although the addition of size specifications or photos with scales theoretically allows for the reconstruction of the ›true‹ proportions, these are not as immediately apparent in the digital form as they are with the analogue object. At repositories the focus is clearly on comparison, on the bigger picture, not on the individuality of the manuscripts (at least at a first glance), even if this material individuality does only come into view and can only be explored by taking the overview as a starting point.

Adjusting the size of a facsimile thus allows for the examination of materiality (e.g. by enlarging a detail of the manuscript page) as well as for its departure from materiality at the same time. The separation of the size-dimensions from the real proportions of the page in virtual space is actually a prerequisite for examining the page when viewing the facsimile and transposing it into different functional contexts. In a similar vein, only the virtual extraction from the codex enables (or at

7 Lechtermann/Stock: Virtuelle Textkonstitutionen, p. 76.

8 Cf. the general definition of models in Stachowiak, Herbert (1973): Allgemeine Modelltheorie, Wien/New York: Springer. In the theoretical discussion of the Digital Humanities, modelling has been highlighted as the key concept of the discipline, McCarty, Willard (2004): »Modeling: A Study in Words and Meanings«, in: Susan Schreibman/John Unsworth/Ray Siemens (eds.): A Companion to Digital Humanities, Oxford: Blackwell, pp. 254–272; Flanders, Julia/Jannidis, Fotis (2018): The Shape of Data in Digital Humanities. Modeling Texts and Text-based Resources, London: Taylor and Francis; Piper, Andrew (2017): »Think Small: On Literary Modeling«, in: PMLA 132, pp. 651–658.

9 This is particularly true for digital synoptic editions: With digital scholarly editions it has become common practice to offer a digital facsimile of the manuscript page next to the transcribed text. This option is also increasingly available with synoptic editions. However, the size of the facsimiles is determined not by the actual size of the codex but rather by the organizing grid layout that arranges the specific windows for each witness: Thus, all facsimiles appear adjusted to the same size. Since synoptic editions usually serve for line-by-line comparisons of differences between versions, the organizing principle of the online display is the text as a transcription, not the text in its material form.

least facilitates) a comparison between different forms of materiality: Manuscript facsimiles can easily be viewed side by side or even analysed with the help of the computer regarding e.g. different layout designs or other palaeographic or codicological peculiarities.[10] The ability to analyse is achieved precisely by distancing oneself from the manuscript object.

2.2 Modelling between material and semiotic text: Markup

As a second variety, contrary to this first effect of digitization, virtualization may also emphasize a notion of text that can be described as semiotic text:[11] In this notion, text is conceived of as a system of characters, abstract from its material and individual concretization. Semiotic texts are produced through the rendering of text in an abstract code, especially in plain text files. However, as has already been seen in the above discussion on the abstractness of facsimiles, it would be too simplistic to map the distinction between material and semiotic text onto that between image and text files. Rather than a strict demarcation, it is a scalable relationship that allows for different degrees of proximity to materiality—or abstractness—when transferring text into code.

This holds true especially since plain text can easily be enriched with markup, as it is the case with XML and the guidelines of the Text Encoding Initiative (TEI).[12] As Patrick Sahle has pointed out, the TEI actually is a hybrid between a material-agnostic abstraction of text and the attempt to retrieve the description of the source document in the encoding.[13] This hybrid status results from a tension between the fundamental purpose of the TEI to provide a data model for all kinds of electronic texts, regardless of their materiality as well as the nature of their source documents, and the tendency of its users to not only record an abstract structure of the text but rather to document the structural properties of the source.

---

10 Lechtermann/Stock: Virtuelle Textkonstitutionen, p. 75. The spatial reorganization can be seen for instance in the scholarly edition »Welscher Gast Digital« (Šimek, Jakub/Schmidt, Peter/Schneider, Christian (eds.) (2015ff.): Thomasin von Zerklaere: Welscher Gast. Text-Bild-Edition ›Welscher Gast digital‹. Heidelberg: Universitätsbibliothek. Online: https://doi.org/10.11588/edition.wgd (last access: 21.09.2025), where all the manuscripts of the transmission that offer pictures or paratext for the same motif of the text can be arranged as a tableau (compare e.g. https://wgd.materiale-textkulturen.de/illustrationen/motiv.php?m=1, last access: 21.09.2025). Thus, instead of the syntagmatic arrangement of the linear text, the paradigmatic relations between the witnesses are emphasized.

11 Lechtermann/Stock: Virtuelle Textkonstitutionen, p. 76

12 Online: https://tei-c.org/ (last access: 25.09.2025).

13 Sahle, Patrick (2013): Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Norderstedt: BoD, volume 3, pp. 341–390, especially p. 355.

As XML, the TEI-scheme follows the basic principles of descriptive markup.[14] Unlike its counterpart, procedural markup, descriptive markup aims to describe data solely in terms of its structure and does not specify how this data should be implemented or presented. Instructions for presentation are reserved for a separate code layer, which can variably be exchanged due to the separation of encoding and presentation.[15] This is what makes descriptive markup powerful and opens up the possibilities for arbitrary transformation that are typical of the digital world: Data stored in descriptive markups can be presented in various ways, whether as online displays or as printed books, simply by changing the code of the representation layer, and it can even be rearranged into indexes or diagrammatic visualizations. Basically, descriptive markup is agnostic with regard to materiality (even though electronic data always has its own materiality as such). Comprehensive functionality can only be achieved by disregarding and abstracting from the material form. Therefore, as was the case with the scalable facsimiles described above, here descriptive markup demonstrates its model properties: only the reduction of details (in this case, the omission of materiality) enables functionality.

However, this logic appears to be weakened with the actual use of the TEI: Precisely because of the shift in editorial practice towards materiality in recent decades, with most projects the TEI does not only serve for describing the structure of an ideal text but also for describing the properties of the source document or the process of text creation.[16] Especially with digital scholarly editions it is now standard practice for all transcriptions as well as reconstructed texts to be encoded in TEI. Most of the editions also trace at least a basic set of the manuscript characteristics like abbreviations, initial letters and corrections.[17] Thus, even if XML-files might be more distant from the material source object than facsimiles, they can still be geared towards the material text.

In the case of digital editions, shifts in the scale from the semiotic back to the material text arise also from their ability to juxtapose any number of text views or variant readings: While single-text editions in the print paradigm were necessarily dependent on reducing this diversity to an abstract edition text even in the case of rich traditions, because more than one text is difficult to display in a printed book,

14 Ibid., pp. 133–156, also for terminological extensions of the dichotomy.

15 Within a web environment, the layer of descriptive markup usually consists of an HTML code, whereas the presentation layer draws on CSS (for layout) and JavaScript (for functionality).

16 This focus, which was only added to the TEI at a later stage, is also evident in the fact that, shortly after its initial formulation, special interest groups (SIGs) were formed for the encoding of manuscripts and genetic editions, the results of which have now been largely integrated into the TEI; P. Sahle: Digitale Editionsformen, vol. 3, pp. 354–355.

17 To a certain extent TEI allows for encoding both, a material and a semiotic text view, by employing the tag <choice>, by which e.g. abbreviations as well as their expansions can be recorded.

the digital medium fosters the presentation of different text versions that are always individual and usually linked to a specific text carrier. Virtualization thus leads to an archive of a variety of text versions that can eventually also be interesting in terms of their materiality.

Of course, as it was the case with the facsimiles, this materiality can only be modelled in the digital realm, i.e. reproduced only approximately, but this modelling leaves its traces in the code. This means that at least in the case of digital editions also the digitization of text as a string of characters (and not only the rendering of text as facsimile images) does not necessarily lead to complete abstraction, as it tends to do with print media, but very often rather to a more meticulous focus on specific details. One might also say that this kind of text encoding offers a greater consideration of complexity compared to print and thus ultimately a complication of the constitution of the text.

2.3 Models for distant reading: Plain text

However, digital editions and their transcriptions are only one way of text encoding. At the other end of the spectrum there is another variant of digitization that is particularly widespread in digital philology, which actually tends even more towards a semiotic constitution of text and is in some respects diametrically opposed to text encoding using markup. I am referring here to the reproduction of texts in plain text files, which is particularly relevant in the field of text analysis.[18]

Digital text analysis usually focuses on large amounts of text, on the macro level. This is at least partly because the most powerful methods of digital text analysis are ultimately statistical and probabilistic in nature. And statistics work particularly well for a large amount of data, not for individual cases.

The best-known approach in this field is distant reading, a concept first introduced in 2000 by Stanford comparative literature scholar Franco Moretti.[19] The starting point for the development of distant reading is provided by Moretti's hard-to-refute observation that the mass of texts that exist in the world is simply too large for any single researcher to read them all. According to a now outdated and, of course, somewhat ironic estimation by Google, fifteen years ago there were exactly 129.864.880 books in the world, a significant portion of which Google wanted to digitize.[20] In contrast, it has been suggested that an avid reader can read a maxi-

18 In some markup typologies these texts are also referred to as ›no markup‹ texts; P. Sahle: Digitale Editionsformen, vol. 3, p. 152.

19 Moretti, Franco (2000): »Conjectures on World Literature«, in: New Left Review 1, pp. 54–68.

20 Cf. Taycher, Leonid (2010): Books of the World, Stand up and Be Counted! All 129,864,880 of you. Online: http://booksearch.blogspot.de/2010/08/books-of-world-stand-up-and-be-counted.html (last access: 21.09.2025).

mum of 10.000 books in their lifetime. This cannot be reconciled; a single person can only ever survey a tiny fraction of the cultural production.

Moretti's answer to this dilemma is well known and quite provocative: since reading more and more books is pointless anyway, Moretti simply recommends stopping reading altogether; or, to put it less bluntly, reading in a different way: instead of close reading, which Moretti considers typical of literary studies, he advocates reading from a distance.[21] Moretti originally conceived of distant reading as a kind of patchwork research in the course of which, instead of the primary texts, only and exclusively secondary research literature is read. But of course, distant reading is particularly suitable when you have a large amount of digitally available texts that can be evaluated by applying digital methods. And one might say that it was only with the transfer of the distant reading approach to the Digital Humanities that the method really took off.

Of course, and Moretti was aware of this, reading from a distance always entails a loss of accuracy—an effect that ultimately always occurs when you look at the big picture, because you have to neglect the subtleties and idiosyncrasies—there is no other way than to accept this loss:

> »If we want to understand the system in its entirety, we must accept losing something. We always pay a price for theoretical knowledge: reality is infinitely rich; concepts are abstract, are poor. But it's precisely this ›poverty‹ that makes it possible to handle them, and therefore to know. This is why less is actually more.«[22]

The necessity for abstraction that Moretti calls for is supported not only by a logical-epistemological argument—if there are more books than one can read, one can never know the whole truth—but also by a moral one: For despite this reading gap, literary studies have of course always been practiced, but only by applying a trick, so to speak. And this trick consists of preselecting the number of texts worthy of study by choosing a canon. Close reading can only be practiced if one considers some books to be better or more valuable than others.[23] However, from a postcolonial perspective this leads to undesirable effects, such as that Western literature by white men is at the heart of the canon. Close reading, one might paraphrase, is therefore not only logically wrong but also, to a certain extent, morally wrong.

At this point at the latest it becomes clear that distant reading not only entails a change of the methodological approach but also of the constitution of the object: the individual text is replaced by a system that must be elucidated »in its entirety«. Instead of dealing with texts in detail, analyses at the macro level neglect nuances.

21 Moretti: Conjectures, p. 57.

22 Ibid., pp. 57–58.

23 Ibid., p. 57.

As a consequence, aesthetic evaluation is less important than the investigation of structural patterns.

In a later essay, which implicitly addresses critical positions towards his distant reading approach, Moretti once again emphasized this shift of focus from individuality to the collective, using Schleiermacher's classical hermeneutics as a foil.[24] While Schleiermacher is concerned with understanding text in its individuality, with formal structural patterns in texts being only a means to an end, quantitative hermeneutics bring about a reversal of these relationships, turning hermeneutics, so to speak, from its head to its feet: »computational criticism clearly reverses the hierarchy: conventions matter for us much more than any individual text: they matter as such, unlike what happens in the hermeneutic project.«[25]

The virtualization of literature in large corpora thus brings along a formalistic approach that is finally justified by Moretti's sociological or even neo-Marxist grounding.[26] For him, formal patterns are primarily relevant in terms of elucidating social power relations: »Forms are the abstract of social relationships: so, formal analysis is in its own modest way an analysis of power.«[27] Unlike Schleiermacher, Moretti is not concerned with understanding individuality and the author better than she understands herself but rather with understanding conventions and formal patterns, »better than their society ever did.«[28] By disregarding the individuality of the texts (and thus ultimately also the material individuality), this third variation of virtualization moves furthest on the scale between ideal and material text towards the pole of abstraction. Its corresponding data format is provided by the markupless plain text, where the materiality of the source document no longer leaves any traces.

What is particularly striking about Moretti's conception is the metaphor of ›entireness‹ and the claim to completeness that seems to be associated with it. In the critical debate surrounding Moretti and the distant reading approaches that followed him, objections to such a claim to completeness soon arose. Jeremy Rosen e.g. contested »that it ought to be a goal of scholarship ›to say with any certainty what

24 Moretti, Franco (2017): »Patterns and Interpretation«, in: Pamphlets of the Stanford Literary Lab 15, pp. 1–10.

25 Ibid., p. 7.

26 Moretti also points out that this results not only in a shift of method but also of the subject of literary studies—from the individual text to a collective, abstract corpus that has no meaning in itself, only formal traces that have been socially attached to it, cf. ibid., p. 1. On the embedding of Moretti's concept in the tradition of the sociology of literature cf. Underwood, Ted (2017): »A Genealogy of Distant Reading«, in: Digital Humanities Quarterly 11. Online: http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html (last access: 21.09.2025).

27 Moretti: Conjectures, p. 66.

28 Moretti: Patterns, p. 8.

contemporary cultural production as a *whole* looks like‹ […], in other words, that analyzing more texts, or even, in a fantastic scenario, everything produced in a given moment, would be necessary or desirable even if it were possible;«[29] And even if Ted Underwood argues that the notion that distant reading entails the construction of a »database that includes ›everything that has been said and thought‹« is but a cliché and that Moretti in his initial conceptions actually was aware that ›entirety‹ can only be approximated by way of samples,[30] I agree with Kathrin Bode that the idea of completeness in the field of distant reading has long been virulent and still leaves its mark.[31]

In a 2016 ›pamphlet‹ from the Stanford literary lab on the relation between the archive and the canon e.g. Mark Algee-Hewitt et al. discern three »preliminary notions«: the published, the archive and the corpus.[32] And they continue:

> »The first is simple: it's the totality of the books that have been published (the plays that have been acted, the poems that have been recited, and so on). This literature that has become ›public‹ is the fundamental horizon of all quantitative work (though of course its borders are fuzzy, and may be expanded to include books written but kept in a drawer, or rejected by publishers, etc.). The archive is for its part that portion of published literature that has been preserved—in libraries and elsewhere—and that is now being increasingly digitized. The corpus, finally, is that portion of the archive that is selected, for one reason or another, in order to pursue a specific research project. The corpus is thus smaller than the archive, which is smaller than the published: like three Russian dolls, fitting neatly into one another.«[33]

The entirety appears graded here and bound to publication, but »the totality of the books« is still the goal, as is the convergence of the three spheres that seems to be in reach with digitization:

---

29 Rosen, Jeremy (2011): »Combining Close and Distant, or, the Utility of Genre Analysis. A Response to Matthew Wilkens's ›Contemporary Fiction by the Numbers‹«, in: Post45. Online: https://post45.org/2011/12/combining-close-and-distant-or-the-utility-of-genre-analysis-a-response-to-matthew-wilkenss-contemporary-fiction-by-the-numbers/ (last access: 21.09.2025).

30 Underwood, Ted (2015): A dataset for distant-reading literature in English, 1700–1922, in: The Stone and the Shell. Online: https://tedunderwood.com/2015/08/07/a-dataset-for-distant-reading-literature-in-english-1700-1922/ (last access: 21.09.2025).

31 Bode, Katherine (2017): »The Equivalence of ›Close‹ and ›Distant‹ Reading. Or, Toward a New Object for Data-Rich Literary History«, in: Modern Language Quarterly 78, pp. 77–106.

32 Algee-Hewitt, Mark et al. (2016): »Canon/Archive. Large-scale Dynamics in the Literary Field«, in: Pamphlets of the Stanford Literary Lab 11, pp. 1–13, here p. 3.

33 Ibid., p. 3.

> »But with digital technology, the relationship between the three layers has changed: the corpus of a project can now easily be (almost) as large as the archive, while the archive is itself becoming – at least for modern times – (almost) as large as all of published literature. When we use the term ›archive‹, what we have in mind is precisely this potential convergence of the three layers into one; into that ›total history of literature‹, to borrow an expression from the Annales, that used to be a mirage, and may soon be reality.«[34]

As the fuzzy borders of books in a drawer or those rejected by publishers might reveal (that also resist the neat fit of the three dolls, since the sphere of the archive can sometimes be larger or simply structured differently from the sphere of the published books), this model of literature appears to be organized around the unit of the printed book, publication processes and other notions connected to the print paradigm. This raises the question of the extent to which it can be transferred to the pre-print conditions underlying medieval literature. To answer this question, it seems advisable to examine the infrastructures that contain digital corpora or the digital archive of medieval literature.

3. Infrastructures for a distant reading of medieval texts

3.1 Counting medieval texts

One of these infrastructures, which holds digital texts for analysis, is the Mittelhochdeutsche Begriffsdatenbank (MHDBDB, Middle High German Conceptual Database), a tremendously worthwhile pioneering project that has existed since the 1970s and offers a somewhat comprehensive collection of medieval German texts.[35] The version of the database that has been online for many years contained exactly 666 texts.[36] The database has been redesigned for several years but this redesign has never progressed beyond beta status. In the course of the relaunch, a few duplicates have been removed, and the beta version now holds 664 entries.[37] In the meanwhile,

34 Ibid, p.3. In the following the authors concede that this is just a theoretical conception that cannot be easily put into practice because the sampling of digital archives is heterogenous. But this does not seem to affect the foundations of the model of the three spheres that converge into one total archive and a ›total history of literature‹.

35 Zeppezauer-Wachauer, Katharina (2022): »50 Jahre Mittelhochdeutsche Begriffsdatenbank (MHDBDB). Eine Jubiläums-Zeitreise zwischen Lochkarten, Pixel-Drachen, relationaler Datenbank und Graphdaten«, in: Elisabeth Lienert et al. (eds.): Digitale Mediävistik. Perspektiven der Digital Humanities für die Altgermanistik. Oldenburg: BmE, pp. 161–86. Online: https://doi.org/10.25619/BME20223203 (last access: 21.09.2025).

36 Online: https://mhdbdb-old.sbg.ac.at/ (last access: 21.09.2025).

37 Online: https://mhdbdb.sbg.ac.at (last access: 21.09.2025).

a Github repository has been published, where the 666 texts of the older version can be retrieved in plain text or in TEI encoding.[38]

These differences in numbers alone reveal that the relationship between the database and Moretti's totality appears to be not a straightforward one. The question that has been bothering me for a long time, therefore, is this one: Is this the system in its entirety of medieval German literature that Moretti speaks of? Is the MHDBDB as complete as Moretti envisions or does it at least come close to representing a population, the convergence of published and archived texts?

Lechtermann and Stock have pointed out that medieval German texts could, in some respects, actually be a very rewarding data corpus because the velocity of the data volume is relatively low.[39] Today, new medieval German texts are simply no longer being produced, except perhaps in the private closets of fanatical philologists. The number of medieval German texts is therefore, in principle, at least limited. However, we will never have all medieval German texts, because some of the texts have been lost in the process of manuscript transmission. Conversely, the limited corpus of medieval German manuscripts could grow precisely because originally lost texts reappear through new manuscript discoveries. But this growth will be marginal.

Thus, one should basically be able to count the number of all medieval German texts. But of course, since medieval German texts are not bound to books, not only the question arises of what is a medieval German text but rather: what is *one* text?

In the context of medieval literature, for example, it is a problem of conceptualization whether each handwritten manifestation of a text should be counted as a separate text or whether the tradition of sometimes more or less divergent text versions should be aggregated into one single work. In the MHDBDB, for instance, four versions of the *Nibelungenlied* are included: The *Nibelungenlied* according to manuscripts A, B and C, but also according to the edition of Bartsch and de Boor, which is based on a mixture of the B and the C text.[40]

The texts of the manuscripts of the *Nibelungenlied* differ in quite a few rather decisive variants, but still there are far ranges of text that are similar. Would it pro-

38 Online: https://github.com/Middle-High-German-Conceptual-Database (last access: 21.09.2025).

39 Cf. Lechtermann/Stock: Virtuelle Philologie, p. 437.

40 However, it seems that this does not result from an increased sensitivity to the medieval transmission but rather, quite pragmatically, from the availability of the manuscript texts in digitized printed editions, as revealed by other comparable cases of texts that are also transmitted in multiple manuscripts but not represented in such a variety in the database. Not least due to the prominence of the *Nibelungenlied*, the text of manuscripts A, B and C have all been published in such print editions. For a reference of the used editions see the text list of the MHDBDB under https://mhdbdb-old.sbg.ac.at/mhdbdb/App?action=TextList (last access: 21.09.2025).

vide a more concise picture of the totality of literature to include each version of the text in a corpus for distant reading, or would that privilege a canonical text over more sparsely transmitted texts? In a frequency analysis of words in texts of the time around 1200 e.g. terms that are typical for the *Nibelungenlied* would be counted four times and skew the results, the same would be true for a descriptive statistic of how many texts were written around that time. But if one includes only one version of the *Nibelungenlied*, which one should it be?

The question of what should be regarded as a (separate) text is also difficult across genres. This is illustrated for example by the pragmatic classification of texts in the MHDBDB: longer novels and chronicles are generally considered to be one text, but short stories are often aggregated into text corpora which are usually organized by author name and counted as one entry in the total of 664 to 666 texts. The situation becomes even more difficult with lyric, because here it is often difficult to say whether different stanzas belong to one song or not. In line with the phenomenon of *mouvance*, the relative instability of verse groups in the tradition, stanzas are sometimes attributed to different songs or even authors in different manuscripts.[41] Unlike in the print era, there are no clear conventions for how verse groups are graphically marked in manuscripts, and the lack of such a clear demarcation is also in accordance with the medieval performance practice, in which stanzas could presumably be freely restructured. The MHDBDB solves this problem in a very pragmatic way, by not recording individual songs but rather aggregating the poetic production and assigning it to an author.

But even with longer texts the equation ›one data record equals one text‹ does not always hold true. For example, the very extensive *Prosa-Lancelot* is divided into three files, obviously primarily for pragmatic reasons. The *Prosa-Lancelot* thus accounts for three of the entries of the MHDBDB. And even for longer texts, the tradition poses pitfalls for assessing the question of text constitution and separation: well-known cases include the formally different texts *Nibelungenlied* and *Nibelungenklage*, which are consistently preserved together in the manuscripts, or the *Mantel*, which replaces the probably lost prologue to Hartmann's of Aue *Erec* in the only complete manuscript.

The second problem with deliminating the ›totality of medieval German literature‹ is that even Middle High German itself is not a clearly defined concept. The MHDBDB contains texts from the 11th to the 16th century, i.e. from a period that encompasses much more than what is considered Middle High German in linguistic history classifications. During this period the German language changed quite significantly, but of course continuously. So, there is no point at which one could say

41 Zumthor, Paul (1984): La Poésie et la voix dans la civilisation médiévale, Paris: Presses universitaires de France.

that some texts are still Middle High German and some texts are not. In addition, the language is not orthographically standardized and varies from region to region.

3.2 Exploration of the database

Leaving aside the question of what a complete collection of Middle High German literature might look like in general, it is quite obvious that, despite its impressive size, the MHDBDB can only offer a selection of medieval German literature. In the terminology of Algee-Hewitt et al. it represents an archive or, so to speak, the second layer of the Russian doll.

Unlike in a library, i.e. a ›real‹ archive, the texts in the virtual database are deprived of their object qualities. E.g. the size of the corpus can initially only be estimated based on the number of database entries, but as explained above, counting the entries is not without problems. If you have texts in book form, you can estimate the length of a text based on the size and thickness of the book. This can sometimes be misleading, as text editions can vary in format and density of printing or include paratext. And, of course, it is even more misleading in the case of medieval literature, since book editions are only derivatives of manuscripts (which in turn are partly derivatives of oral practices). But if you disregard this inaccuracy, you get a rough idea of the length of the text in analog form. It is of course unlikely that a single researcher has seen all medieval German texts at one library or even anywhere near as many as there are in the MHDBDB. But it is very likely that she has seen individual texts side by side and already has a sense of the proportions. With the digital repository it is almost the opposite: Theoretically, you have all the texts at a glance, but the entries reveal nothing about the length of the text. The file size would provide a rough indication or, more precisely, the number of words contained in a text file. However, simply listing the numbers would not provide clarity with such a large corpus, since the number of data points is simply too large; only the presentation of the numerical ratios in a diagram provides an intuitive ›feel‹: In the virtual world, diagrammatic visualizations serve as surrogates for lost sensory quality.

Fig. 1: Word count of all texts of the MHDBDB in chronological order.

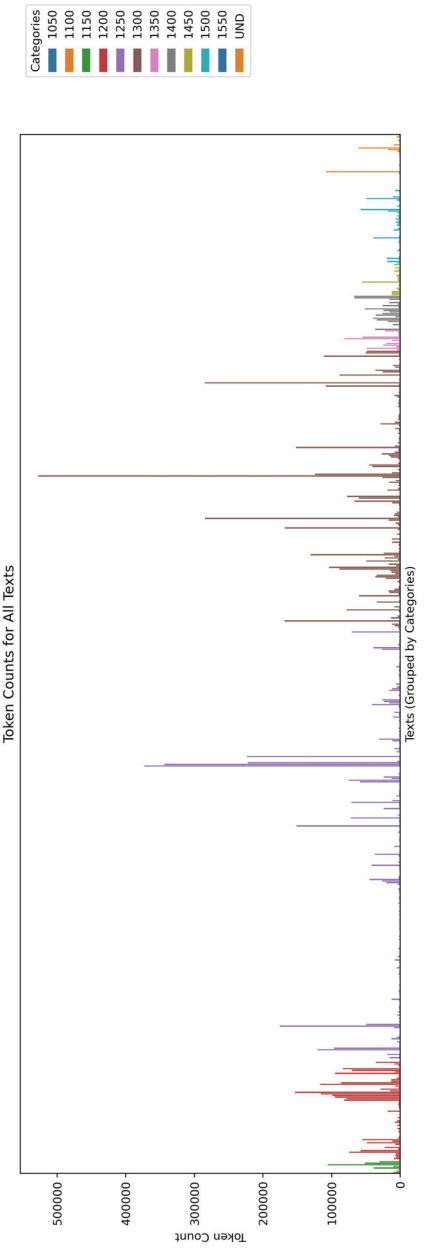


Figure 1 shows a barplot of the word count of all the texts in the database. The texts are distributed on the x-axis according to temporal bins, covering timespans of 50 years. Thus, the x-axis builds an approximate timeline, with older texts to the left and more recent texts tending to the right, while the height of the bar indicates the text size in tokens. However, a chronological division presents the next problem with classifying the corpus: In the case of medieval German literature, such a division can once again only be done imprecisely, because a large part of the pre-modern texts are not provided with an exact date in their transmitting manuscripts, but this can only be deduced secondarily, for example from references and allusions in the text. This means that many of these data are also controversial in research. Nevertheless, I have attempted to divide the texts into approximately 50-year periods, which I have compared with relevant encyclopedias such as the *Verfasserlexikon*.[42] Texts that offer no clues at all as to their chronological classification have been placed in a separate group (UND, for undefined). These are shown on the far right of the diagram.

What immediately catches the eye with the diagram is the fact that the length of the texts varies greatly. Texts with only a few tokens are juxtaposed with those with up to more than 500.000 tokens.

Fig. 2: Distribution of the Token count in the MHDBDB

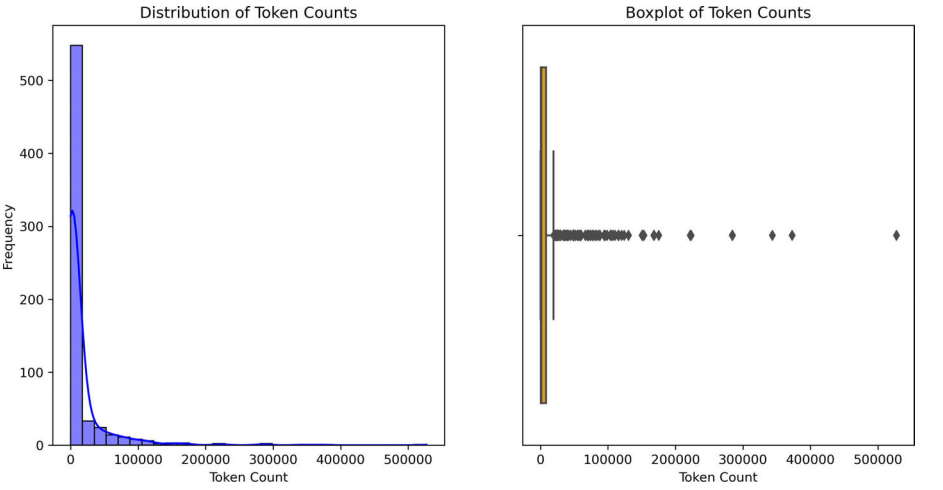


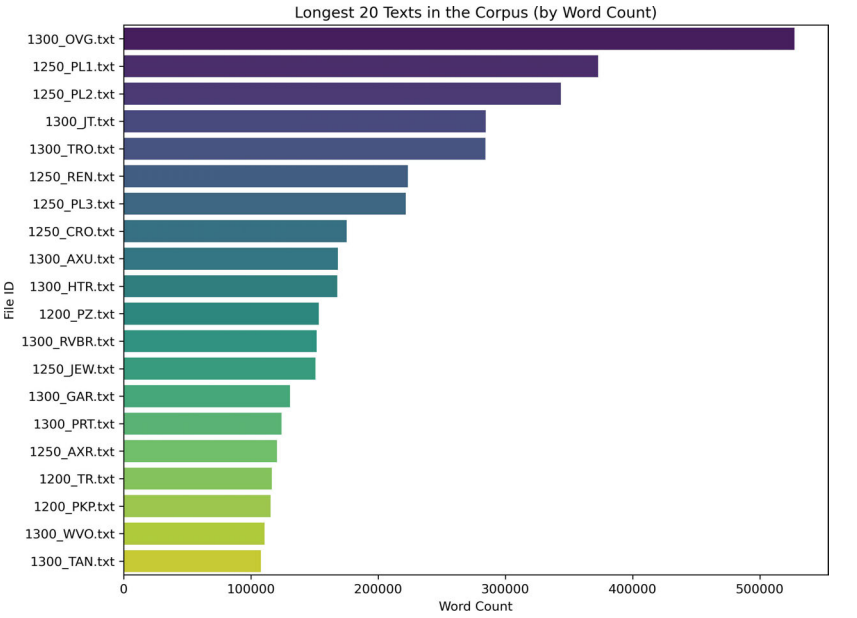
Figure 2 provides two plots that offer a more detailed picture of the distribution. As can be seen on the left, a vast majority of 548 texts belong to the first bin

42 Ruh, Kurt et al. (eds.) (1978ff.): Die deutsche Literatur des Mittelalters. Verfasserlexikon. 2nd, revised edition, Berlin/New York: De Gruyter.

that ranges from 51 tokens (the size of the smallest text in the corpus) to 17.625 tokens. A similar picture can be seen in the boxplot. The average count for all texts is 15.172 tokens, but deviations are extremely high, ranging up to 527.245 tokens for the longest text.

Figure 3 shows the 20 most extensive texts of the corpus.[43] Rank 1 is Ottokar's *Steirische Reimchronik*, but positions 2, 3 and 7 are held by the above-mentioned parts of the *Prosa-Lancelot* that was split up in the MHDBDB. If all three parts were combined, the text would encompass 938.283 tokens and stand out from the other texts even more than the *Reimchronik*.

Fig. 3: The 20 most extensive texts of the corpus

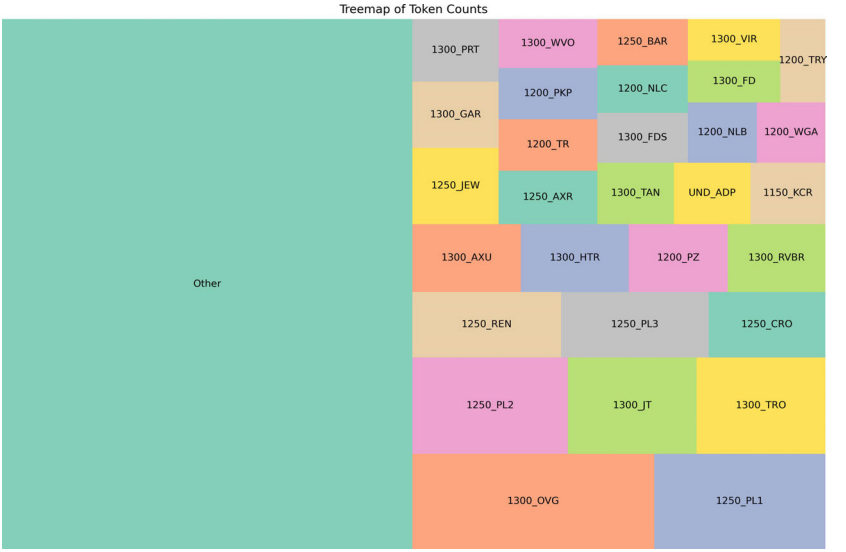


Finally, the treemap in Figure 4 shows how much the proportions are distorted by just a few texts, revealing that the first 30 texts account for half of the entire corpus.[44]

43 The file names are drawn from the short references of the MHDBDB, preceded by the first year of the time bin they have been sorted into. The resolution of these abbreviations (which are not relevant in my context) can be done by help of the MHDBDB text list at https://mhdbdb-old.sbg.ac.at/mhdbdb/App?action=TextList (last access: 21.09.2025).

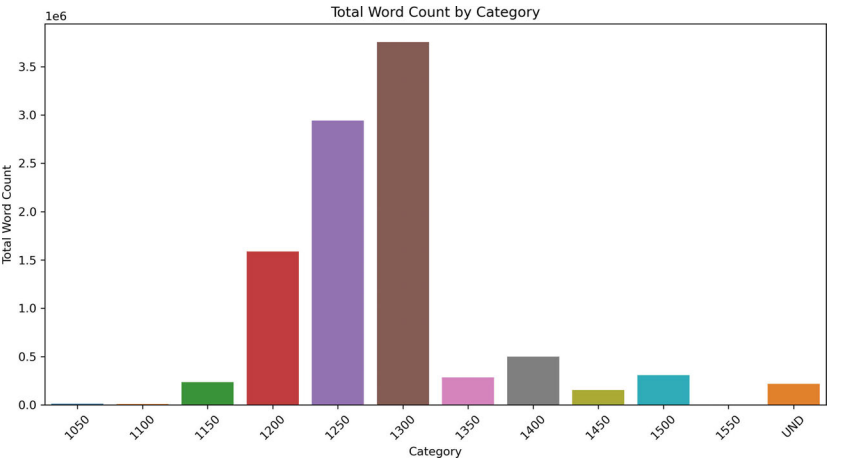
44 The diagram visualizes the single texts as rectangles that are sized according to the token count. As can be seen, two of the versions of the *Nibelungenlied* mentioned above are also among these thirty texts (NLB and NLC), which makes the problem of their multiple versions particularly acute.

Fig. 4: Treemap of size relations in the corpus.



However, this is not the only imbalance of the corpus. If we aggregate the texts from the individual periods over 50 years, it becomes apparent that some of these periods are very poorly represented, whereas others make up a large part of the corpus, as can be seen in Figure 5.

Fig. 5: Number of Tokens by 50-year periods.



As can be seen, the MHDBDB has a clear focus on the period from 1200 to 1350.[45] However, it cannot be expected that the balance of the corpus will improve significantly by adding more and more texts. For some periods there are simply not enough texts available to close the gap. In addition, the problem of the great heterogeneity of the corpus would persist or even increase, as it is not possible to ensure uniformity when adding new texts.

Thus, if one wants to examine the ›system in its entirety‹ by drawing on more and more texts, one will soon notice that necessarily this system shows a degree of imbalance and is dominated by certain texts and eras. This becomes particularly clear when we have a relatively small corpus at our disposal, such as medieval German.

Such a distortion of the archive does not mean, of course, that analyses based on the material are fundamentally impossible. What is required is a skilful selection of the corpus from the archive, according to the model of Algee Hewitt et al., the transition from the second to the third layer.[46] However, any selection leads to a restriction of the data material which may cause problems in languages such as Middle High German, where the amount of data is ultimately limited.

These problems are likely to become particularly relevant when looking at the latest developments in natural language processing towards large language models. The decisive breakthroughs of recent years, which have enabled computers to generate meaningful texts or translate into other languages, for example, are ultimately based on the shift in computational linguistics toward distributional semantics and the representation of words in word embeddings. Distributional semantics is based on the assumption that the meaning of a word does not arise from the word itself but from the context within which it appears.[47] For example, if the word ›bank‹ is surrounded by many expressions from the financial sector, then it is likely to refer to the financial institution. If there are many words referring to nature in the vicinity of the word, then the meaning is likely to be that of a landscape feature.

These contextual probabilities can also be expressed as numerical values and packed into vectors, i.e. series of numbers that characterize words as embeddings and can also be used for calculations. This allows for a surprisingly good modeling

45 This may be due to the original focus on ›Middle High German‹ evident in the title of the database, as this would be the period for which this language stage can actually be roughly estimated.

46 In corpus linguistics the problem of distortion is counteracted by creating balanced corpora. One such balanced corpus is the Middle High German Reference Corpus, currently the second largest resource for medieval German literature (Online: https://www.linguistics.rub.de/rem/index.html, last access 21.09.2025). However, the necessary reduction means that the corpus is significantly smaller than the MHDBDB (2 million tokens compared to approx. 10 million), and not all texts of literary significance can be included.

47 Cf. Lenci, Allesandro/Sahlgren, Magnuns (2023): Distributional Semantics. Cambridge et al.: Cambridge University Press.

of deeper text meanings that go beyond the explicit surface of the text. However, to calculate these context probabilities with any degree of stability, a great deal of data is required.

Even though modern language models offer the possibility of transfer learning and the fine-tuning of existing models, the amount of data plays a decisive role, which is why one would not want to further limit the text basis. A corpus philology that makes use of these methods therefore cannot always afford to constrain itself to certain domains (like genres or epochs) that are too finely subdivided. With the embedding model, texts of all kinds are mixed together; who wrote them is irrelevant, they lose their author signature and become training material of equal rank. This represents the highest level of abstraction: not only do these texts lose their connection to a specific, individual carrier, but as semiotic texts they are no longer relevant in their individuality, but only as components of the corpus.

However, this circumstance seems to be much more precarious for a text corpus such as medieval German, for which only small amounts of text and thus training data are available, than for modern languages, where much more extensive data sets can be drawn upon. I will attempt to demonstrate this by referring to a few experiments with word embeddings that I have conducted for Middle High German. Specifically, I trained a token-based embedding model on the MHDBDB corpus using the Word2Vec method. Word2Vec is a simplified method for creating word embeddings that was developed by Google in 2013.[48]

Even training such a model, which is still quite basic compared to today's LLMs, requires extensive data sets, which raises the question of whether what we have in the MHDBDB, for example, is sufficient for creating such a model in a valid way. The database encompasses approximately 10.000.000 tokens, which is not insignificant, but it must be borne in mind that, as already mentioned, the MHDBDB contains a wide variety of texts, some of which represent different levels of German and may exhibit dialectal peculiarities. The training material is therefore by no means as uniform as it would be in the case of modern languages.

Nevertheless, at a first glance it seems possible to generate a reasonably useful model.[49] Once a Word2Vec model is created, one can approximate the similarity of words by comparing their embedding vectors, and also perform calculations with

48 Cf. Mikolov, Tomas et al. (2013): »Efficient Estimation of Word Representations in Vector Space«, in: 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4. Workshop Track Proceedings, pp. 1–12. Online: https://arxiv.org/pdf/1301.3781.pdf (last access: 21.09.2025). See for an introduction also Schumacher, Mareike (2023): »word2vec«, in: forTEXT. Literatur digital erforschen. Online: https://fortext.net/routinen/methoden/word2vec-1 (last access: 21.09.2025).

49 Cf. Viehhauser, Gabriel (2024): »Digitale Methoden für die Altgermanistik. Eine exemplarische Bestandsaufnahme«, in: Literaturwissenschaftliches Jahrbuch 65, pp. 9–60. Online: https://doi.org/10.3790/ljb.2024.1445301 (last access: 21.09.2025).

these vectors. In most models of natural language texts, subtracting the vector of the word ›man‹ from the vector of the word ›king‹ and adding the vector of the word ›woman‹ results in a vector that is closest to the vector of the word ›queen‹. Such a calculation obviously also works for a model trained on the MHDBDB. Table 1 lists the ten words that show the highest vector similarity to the result of the calculation *küninc + vrouwe – man*, and indeed, the vector of *küninginne* is in first place.

Table 1: most similar words to the vector calculation küninc + vrouwe – man

küninginne	0.64
juncvrouwe	0.58
tohter	0.53
küningîn	0.51
Gezogenlîche	0.48
Herzogîn	0.46
Maget	0.46
Keiserinne	0.44
Neve	0.44
Swester	0.43

However, further exploration of the model has revealed that the embeddings only work well for words that occur with reasonably high frequency in the already relatively small and heterogeneous corpus. The representations for words that occur less frequently are far less stable. I therefore endeavoured to expand the corpus with additional digital texts and trained new models on this expanded corpus. However, the surprising effect that occurred was that at a first glance, some of the models did not appear to have improved but rather had become worse. For example, after adding some texts from Digitales Mittelhochdeutsches Textarchiv,[50] the Word2Vec model was no longer able to ›correctly‹ calculate the equation ›king‹ minus ›man‹ plus ›woman‹ because the word ›juncvrowe‹ had overtaken the word ›küniginne‹ in the ranking.

At the same time, however, the question immediately arises: Can one say that this model is wrong at all? Or does it simply reflect the semantic relationships that come to the fore when less canonical texts are fed into the corpus? Perhaps in the semantic reality of these texts, ›king‹ does not behave in relation to ›man‹ in the

50 Online: http://www.mhgta.uni-trier.de/ (last accessed 21.09.2025).

same way as ›queen‹ does in relation to ›woman‹, but rather in the same way as ›juncvrouwe‹ does in relation to ›woman‹. Once again, the question arises of what the ›totality of texts‹ would look like for medieval German texts: The MHDBDB tends to focus on literature in the emphatic sense, especially courtly literature of the 13th and 14th centuries. A logical extension of the textual material could be, for example, the inclusion of more religious texts (e.g. legends of saints, which are only comparatively sparsely represented in the database) or even *Gebrauchstexte* (such as charters, that are available in large numbers at the monasterium.net archive),[51] but it seems quite clear that such a shift of the domain would lead to totally different results in the conception and analysis of a possible ›system in its entirety‹. At least for medieval texts, the »preliminary notion« of the totality of »published literature« seems to be not as simple, as Algee-Hewitt et al. suggest.

4. Conclusion: On the adequacy of virtual models

As has been seen, the virtualization of medieval texts can lead to a wide range of effects that can be sorted along a scale that ranges from notions of a more concrete, material text to a more abstract, ideal text. Virtual reconstructions always—and necessarily—miss their subject, they are models of texts rather than these texts themselves. Even facsimiles, which are photographically exact reproductions of the material text carrier, are not limited to the depiction of the source document but are entities with their own functionality precisely because of their difference from this text carrier. The same holds true in more evident cases such as the conception of texts in large corpora. Here, the license to abstraction and to neglect details is particularly strong. In the macro perspective, the individual text is not important or at least not important in all its details.

Nevertheless, it seems crucial to me not to overlook the fact that these virtualizations, despite their oblivion of detail, also have a strong claim to tell the truth about literature, in a similar way as a facsimile brings with it the promise of representing the original text of the manuscript. Lechtermann and Stock have pointed out that behind every editorial reproduction of a medieval text there is a promise of authenticity which consists of claiming to reproduce the text correctly, even though this reproduction does not take place in the original form of the manuscript.[52] The edited text is *the* text or should be the text itself, even if it was originally transmitted differently.

51 Online: https://www.icar-us.eu/cooperation/online-portals/monasterium-net/ (last accessed 21.09.2025).

52 Lechtermann/Stock: Virtuelle Textkonstitutionen, p. 69.

In the case of plain texts in text corpora it seems to me that a promise of authenticity also comes into play, albeit in a less explicit form than is the case with edited texts. The digital texts from the MHDBDB do not blatantly proclaim their authenticity. In fact, some of the text editions the database builds on are from editions that would now be considered outdated and no longer authentic. On the contrary, the compilation of texts is primarily pragmatic in nature, as the operators of the database would readily admit.

When conducting text mining on these texts, complete authenticity is not essential. As explained above, a macro-perspective, distant reading analysis is not overly concerned with detail anyway, so a degree of vagueness can be tolerated; the focus is on the big picture. As the example of the *Nibelungenlied* has shown, taking into account nuances in the history of transmission and text variants can even lead to difficulties that might distort the statistical findings.

Thus, the fact that macro analyses do not take into account all the details cannot be blamed on them. In many cases it is precisely the distance of the model that produces a clearer picture.[53] The more critical question would rather be if the model is actually appropriate for the domain on the one hand and whether the modelling does not unconsciously make claims to authenticity that are not clarified.

In this regard, it seems to me that empirical evaluation brings a claim to truth or authenticity back into play which culminates in Moretti's idea of »understanding the system in its entirety« (see above): only if you have all the texts, or at least large quantities of them, can you finally learn the truth about literature, and in particular about literature that does not belong to the canon. As I have tried to show, this notion seems to be driven inherently by a book paradigm and therefore is not easily applicable to medieval literature, because of the specific textuality of its texts, but also because of the limited scope of the medieval corpus that fray into different domains that are not easy to set off against each other. The greatest infrastructural challenge would be to combine the extremely abstract textual concept of digital corpora with the concrete handwritten manifestations of text versions of medieval texts.

Bibliography

Algee-Hewitt, Mark/Allison, Sarah/Gemma, Marissa/Heuser, Ryan/Moretti, Franco/Walser, Hannah (2016): »Canon/Archive. Large-scale Dynamics in the Literary Field«, in: Pamphlets of the Stanford Literary Lab 11, pp. 1–13.

53 Cf. in a similar vein Underwood, Ted (2016): »The real problem with distant reading«, in: The stone and the shell. Online: https://tedunderwood.com/2016/05/29/the-real-problem-with-distant-reading/ (last access: 21.09.2025).

Baisch, Martin (2013): »Textualität – Materialität – Materialität – Textualität. Zugänge zum mittelalterlichen Text«, in: Literaturwissenschaftliches Jahrbuch 54, pp. 9–30.

Bode, Katherine (2017): »The Equivalence of ›Close‹ and ›Distant‹ Reading. Or, Toward a New Object for Data-Rich Literary History«, in: Modern Language Quarterly 78, pp. 77–106.

Da, Nan Z. (2019): »The Computational Case against Computational Literary Studies«, in: Critical Inquiry 45, pp. 601–639.

Flanders, Julia/Jannidis, Fotis (eds.) (2018): The Shape of Data in Digital Humanities. Modeling Texts and Text-based Resources, London: Taylor and Francis.

Lechtermann, Christina/Stock, Markus (2020): »Virtuelle Philologie«, in: Dawid Kasprowicz/Stefan Rieger (eds.): Handbuch Virtualität, Wiesbaden: Springer, pp. 425–454.

Lechtermann, Christina/Stock, Markus (2021): »Virtuelle Textkonstitutionen: Die Philologie und ihre mittelalterlichen Objekte«, in: Stefan Rieger/Armin Schäfer/Anna Tuschling (eds.): Virtuelle Lebenswelten: Körper – Räume – Affekte, Berlin/Boston: De Gruyter, pp. 63–86.

Lenci, Allesandro/Sahlgren, Magnuns (2023): Distributional Semantics. Cambridge et al.: Cambridge University Press.

Limpinsel, Mirco (2016): »Was bedeutet die Digitalisierung für den Gegenstand der Literaturwissenschaft?«, in: Zeitschrift für digitale Geisteswissenschaften. Online: https://zfdg.de/2016_009 (last access: 21.09.2025).

McCarty, Willard (2004): »Modeling: A Study in Words and Meanings«, in: Susan Schreibman/John Unsworth/Ray Siemens (eds.): A Companion to Digital Humanities, Oxford: Blackwell, pp. 254–272.

Mikolov, Tomas/Chen, Kai/Corrado, Greg/Dean, Jeffrey (2013): »Efficient Estimation of Word Representations in Vector Space«, in: 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4. Workshop Track Proceedings, pp. 1–12. Online: https://arxiv.org/pdf/1301.3781.pdf (last access: 21.09.2025).

Moretti, Franco (2000): »Conjectures on World Literature«, in: New Left Review 1, pp. 54–68.

Moretti, Franco (2017): »Patterns and Interpretation«, in: Pamphlets of the Stanford Literary Lab 15, pp. 1–10.

Piper, Andrew (2017): »Think Small: On Literary Modeling«, in: PMLA 132, pp. 651–658.

Reiter, Nils/Pichler, Axel/Kuhn, Jonas (eds.) (2020): Reflektierte algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt, Berlin/Boston: De Gruyter.

Rosen, Jeremy (2011): »Combining Close and Distant, or, the Utility of Genre Analysis. A Response to Matthew Wilkens's ›Contemporary Fiction by the Numbers‹«,

in: Post45. Online: https://post45.org/2011/12/combining-close-and-distant-or-the-utility-of-genre-analysis-a-response-to-matthew-wilkenss-contemporary-fiction-by-the-numbers/ (last access: 21.09.2025).

Ruh, Kurt/Keil, Gundolf/Schröder, Werner/Wachinger, Burghart/Worstbrock, Franz Josef (eds.) (1978ff.): Die deutsche Literatur des Mittelalters. Verfasserlexikon, 2. völlig neu bearbeitete Auflage, Berlin/New York: De Gruyter.

Sahle, Patrick (2013): Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Norderstedt: BoD.

Schumacher, Mareike (2023): »word2vec«, in: forTEXT. Literatur digital erforschen. Online: https://fortext.net/routinen/methoden/word2vec-1 (last access: 21.09.2025).

Šimek, Jakub/Schmidt, Peter/Schneider, Christian (eds.) (2015ff.): Thomasin von Zerklaere: Welscher Gast. Text-Bild-Edition ›Welscher Gast digital‹. Heidelberg: Universitätsbibliothek. Online: https://doi.org/10.11588/edition.wgd (last access: 21.09.2025).

Stachowiak, Herbert (1973): Allgemeine Modelltheorie, Wien/New York: Springer.

Strohschneider, Peter (1997): »Situationen des Textes: Okkasionelle Bemerkungen zur ›New Philology‹«, in: Helmut Tervooren/Horst Wenzel (eds): Philologie als Textwissenschaft. Alte und neue Horizonte, Berlin: Schmidt, pp. 62–86.

Strohschneider, Peter (1999): »Textualität der mittelalterlichen Literatur: Eine Problemskizze am Beispiel des ›Wartburgkrieges‹«, in: Jan-Dirk Müller/Horst Wenzel (eds.): Mittelalter: Neue Wege durch einen alten Kontinent, Stuttgart/Leipzig: Hirzel, p. 19–41.

Strohschneider, Peter (2014): Höfische Textgeschichten. Über Selbstentwürfe vormoderner Literatur, Heidelberg: Winter.

Taycher, Leonid (2010): Books of the World, Stand up and Be Counted! All 129,864,880
of you. Online: http://booksearch.blogspot.de/2010/08/books-of-world-stand-up-and-be-counted.html (last access: 21.09.2025).

Underwood, Ted (2015): A dataset for distant-reading literature in English, 1700–1922, in: The Stone and the Shell. Online: https://tedunderwood.com/2015/08/07/a-dataset-for-distant-reading-literature-in-english-1700-1922/ (last access: 21.09.2025).

Underwood, Ted (2016): »The real problem with distant reading«, in: The stone and the shell. Online: https://tedunderwood.com/2016/05/29/the-real-problem-with-distant-reading/ (last access: 21.09.2025).

Underwood, Ted (2017): »A Genealogy of Distant Reading«, in: Digital Humanities Quarterly 11. Online: http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html (last access: 21.09.2025).

Viehhauser, Gabriel (2024): »Digitale Methoden für die Altgermanistik. Eine exemplarische Bestandsaufnahme«, in: Literaturwissenschaftliches Jahrbuch

65, pp. 9–60. Online: https://doi.org/10.3790/ljb.2024.1445301 (last access: 21.09.2025).

Zeppezauer-Wachauer, Katharina (2022): »50 Jahre Mittelhochdeutsche Begriffsdatenbank (MHDBDB). Eine Jubiläums-Zeitreise zwischen Lochkarten, Pixel-Drachen, relationaler Datenbank und Graphdaten«, in: Elisabeth Lienert/Joachim Hamm/Albrecht Hausmann/Gabriel Viehhauser (eds.): Digitale Mediävistik. Perspektiven der Digital Humanities für die Altgermanistik. Oldenburg: BmE, pp. 161–86. Online: https://doi.org/10.25619/BME20223203 (last access: 21.09.2025).

Zumthor, Paul (1984): La Poésie et la voix dans la civilisation médiévale, Paris: Presses universitaires de France.

List of Figures