ANDREW PIPER AND JAMES MANALAD

Measuring Unreading

When Margaret Cohen first coined the term "the great unread," she was referring to the vast swaths of literary history that remain unexamined by scholars.[1] In this sense, these books are not "not read," but simply not studied, or studied only by a small cohort of later readers. Indeed, many of these supposedly unread books may have been some of the most popular, and thus most read, by readers of the past. As Franco Moretti's later work would underscore, addressing the great unread is about reclaiming the study of certain types of books that have been overlooked and attempting to construct a more representative sample of the past.[2] New digital techniques of text analysis have allowed scholars to be less selective about the books they study and have allowed them to reconstruct more accurate representations of past literary practices.[3]

Left out of this discussion is the larger question of what it means for something to be "unread." There are many books, for example, that we have read but no longer remember anything about. Do these count as read or unread? There are also many books (all too many) that we stopped reading well before the end. Our bookshelves are littered with books with bookmarks positioned far from the back cover. Do these count as read or unread? And there are numerous moments when we've read a passage but stopped paying attention. Our attention waxes and wanes when we read. What is read and unread in such scenarios?

Historians of reading have long wrestled with these problems. What can we know about what people have read? As Leah Price has suggested, when a book most impresses us, we are often apt to put the pen down.[4] When a book marks us, we don't necessarily mark it. Studying marks, whether in the form of underlining, marginal annotation (or now highlighting), misses a good deal about what matters to reading. Part of the value of reading has been its inscrutability, its capacity as a cultural practice to wall us off from the world.[5]

Reading and unreading might be thought of then as complementary practices, as behaviors that are intertwined with each other. Rather than think about the unread as something distinct and apart, we can also think about it as an integral practice to how we read. In this essay, we want to introduce a new computational method that we can use to think about this imbricated practice of unreading. Where much of the discussion about the great unread has focused on books (or documents more generally), here we want

to explore the differential spaces of reading and not reading *within* books. The great unread is typically thought of as a sampling problem, where the goal is to add more texts to the study of the past. The explicit aim is to move past the study of a single canonical figure, which animates publications like the *Goethe Jahrbuch*. In one sense, then, focusing on the great unread in a journal devoted to a single person my seem paradoxical.

But this presupposes that attention to Goethe's own writing—our reading of his writing—is largely uniform, that there are no unread spaces in Goethe's corpus. Literary criticism is often animated by a spirit of rereading—revisiting a passage or work that others have already commented on and adding a new insight or version of its meaning. In this sense, literary criticism lavishes attention where attention has already been paid. It is a highly unequal enterprise by nature. Our aim in drawing attention to the unread spaces of Goethe's corpus is to initiate a conversation about the meaning of concentration and repetition in the practice of critical reading. What are the values associated with the intense levels of rereading in our field? And what might the value be of a form of criticism that is attentive to the spaces of inattention—that is, to the act of reading the unread, even within spaces that are considered to be highly read (like Goethe)?

We will be focusing on the practice of quotation by Goethe scholars in order to understand the explicit spaces of reading and not reading within Goethe's corpus. *Quotation* for us refers to the scholarly practice of reproducing the words of another writer in the body of an article, which is different from the practice of *citation*, where the emphasis is on referencing a work usually through some indexical typographic marker like a footnote.[6] As we indicate above, the absence of a quotation does not necessarily indicate the absence of reading. But it does offer a lens into cultural and professional valuation—which words of Goethe's do we most often reproduce in our articles and which are most often left out? While we cannot know if unquoted passages or works have not been read, we can learn which ones do not warrant the close critical attention of scholars. These absences mark the absence of a particular kind of reading.

The primary tool that we will be using falls under the heading of a "text reuse algorithm."[7] It allows us to estimate passages in a source corpus (in this case Goethe) that have been reused or "quoted" in a target corpus. For the purposes of this essay, we will be using 68 volumes of the *Goethe Jahrbuch* that have been digitized as the target corpus and the digitized edition of the Weimarer Ausgabe of Goethe's works as the source corpus (not including Goethe's letters).[8] In terms of the algorithm, we use Jonathan Reeves's implementation that is freely available on Github.[9]

When trying to study text reuse, there are many things to consider. The two main challenges are identifying repeated sequences of words that are not quotations, but rather common phrases that simply belong to the language. How can we be certain that a sequence of words belongs to Goethe and is not just a property of speaking German? The second problem has to do with both human and machine error. When critics quote authors, they don't always do so correctly. Words can be dropped or word order can be slightly altered or punctuation might be missing. Machines, too, make mistakes

when, for example, printed volumes are digitized, as has been the case with our data. Some words are mistranscribed and thus will not match.

Text-reuse algorithms provide a variety of ways to address these problems, such as focusing on a minimum number of words in a sequence to match or using methods of fuzzy matching, such as "Levenshtein distance," which allows for the partial matching of words. And yet no matter what we find, we will always be using an estimation when it comes to quotation, not a true number. While this might cause scholars to turn away from these methods, it is important to point out that there is no way to address large-scale questions like the "great" unread without introducing uncertainty into our understanding of the problem.

In order to give readers an idea of how the matching process works before we move on to our analysis, we provide an example of what a match can look like, drawn from a line from the dramatic fragment of *Prometheus*. The highlighted text is what the algorithm has matched on and the surrounding text is provided by the model to give context to the quotation for the purposes of validation. All of the extracted quotations are available for view in our accompanying data and code.

> ***Goethe Jahrbuch*, vol. 1**
> Göttern das. Den unendlichen. Pr Göttern? Ich binn kein Gott 35 Und bilde mir fo viel ein als einer. Unendlich! Allmächtig! Was könnt ihr Sprössling. 23 fie. 27 trotzen. 28—30 ist, nachdem

> ***Prometheus* (Dramatisches Fragment)**
> Göttern das, 32 Den Unendlichen? Prometheus Göttern? Ich bin kein Gott, 34 Und bilde mir so viel ein als einer. 35 Unendlich?—Allmächtig?—36 Was könnt Ihr Könnt Ihr den weiten Raum

Notice how optical character recognition (OCR) errors like "binn" and "fo" are still able to be matched due to the fuzzy matching technique, just as the differing punctuation—the question marks in the original versus the exclamation points in the *Goethe Jahrbuch*—are similarly resolved. But it is also important to point out that this tool does not capture the entirety of the quoted line. The quotation extends further backward and forward in ways that our algorithm does not capture. In other words, we cannot use it to measure the exact number of quoted words. However, it does an excellent job of identifying when a sequence of words from Goethe's corpus is being quoted and attributing that quotation to one or more source texts.

Forms of Attention in Goethe's Corpus

Applying the algorithm described above, we estimate the number of quotations found in the *Goethe Jahrbuch* that refer to various works in Goethe's corpus. Doing so allows us to infer insights about scholarly forms of attention and valuation—not only where that attention lies (i.e., which of Goethe's works are most often quoted) but also the nature of that attention (i.e., what are the semantic qualities associated with the quoted language). For example,

Table 1. List of the top ten most-quoted works from Goethe's corpus in the *Goethe Jahrbuch*, as well as a selection of the unquoted work.

Most Quoted Works	Quotations	Unquoted Works
Faust	624	*Antheil an Lavaters Physiognomischen Fragmenten*
Dichtung und Wahrheit	480	*Die ungleichen Hausgenossen*
Italienische Reise	417	*Lili*
West-östlicher Divan	347	*Benvenuto Cellini 2. Teil*
Farbenlehre	283	*Beiträge zur Optik 2. Stück*
Wilhelm Meisters Lehrjahre	280	*Die vereitelten Ränke*
Allgemeine Naturlehre	215	*Nachspiel zu Ifflands Hagestolzen*
Schriften zur Literatur	192	*Der Schutzgeist*
Maximen und Reflexionen	179	*Balladen* (Late)
Zur Morphologie	172	*Idyllen* (Late)

we can see how the frames of attention in the *Goethe Jahrbuch* adhere to at least some expectations concerning the canon of Goethe's works (see Table 1). *Faust* is the most quoted work overall (split evenly between the first and second parts), followed by Goethe's autobiographical work, and then his scientific writing. The *West-östlicher Divan (West-Easterly Divan*) ranks higher than I would have expected as are the critical essays on literature. Part of the *Divan's* elevation is due to the "Notes" or *Abhandlungen* which is quoted almost as much as the poetry.

While this conforms to expectations it is important to point out that we did not "know" this already in any concrete sense. If we had polled Goethe scholars in advance, while most would have guessed *Faust* would be squarely in the lead, the rest of the list would have been a scattershot of guesses. Similarly, the relative ratio of quotation would have been a total mystery, and we suspect so too would an estimation of the bottom end of the tail—the unquoted or minimally quoted works. Such empirical work has the advantage of giving us much more precise estimates for certain questions. It can also illustrate the severity of concentration, which we can then compare to other authors or fields (Fig. 1).[10] In Figure 1 we can see how evenly or unevenly distributed the attention to certain works is across the collection of essays in the *Goethe Jahrbuch*. The x-axis ranks Goethe's works by the number of quotations that refer to them while the y-axis estimates how many quotations can be attributed to each work. (For example, the top two points represent *Faust I* and *II*.) Like much scholarly behavior, this distribution of attention exhibits a high degree of concentration around a few leading works. It is defined by a basic inequality.

Readers will note that other than the *Divan*, poetry appears to be less central to this list. However, this may be due to the fact that Goethe's poetry is presented across multiple volumes in the Weimarer Ausgabe (or because there is simply less of it). To understand the differential attention across genres in Goethe's writing, we can also focus on the different genres (as we

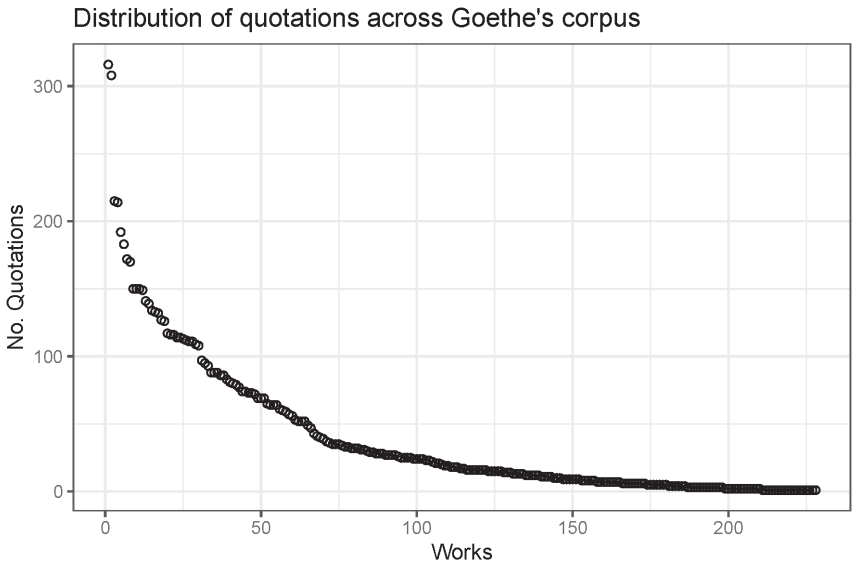

Fig. 1. The quotation counts by works in Goethe's corpus. The x-axis ranks Goethe's works by the number of quotations in the *Goethe Jahrbuch*, while the y-axis estimates the number of quotations per work. The two high points on the left represent *Faust I* and *II*.

have annotated them) instead of the works (see Table 2). Doing so, we see that prose (fiction and autobiographical writing) consumes the most amount of attention in the *Goethe Jahrbuch* and natural scientific writing the least. Poetry is still the second-least quoted genre in Goethe's corpus.

And yet, if we normalize by the overall volume of text—that is, if we calculate how many citations we see per 100,000 words (the length of a standard novel)—poetry now appears to be quoted much more frequently than we might expect, while Goethe's scientific writing is still quoted considerably less frequently. While Goethe's prose fiction and autobiographical writings dominate the most attention, relative to how much there is to be quoted, his poetry is quoted more often.

In addition to genre, we can also get a sense of the temporal focus on Goethe's writing—Are all periods of his life treated equally? Doing so, we see how the late period is overwhelmingly the focus. Based on the above this should not surprise us—*Faust*, *Divan*, *Dichtung und Wahrheit* (*Poetry and Truth*), and much of the morphological writings are all from the late period. The caveats are that the dating we are using is dependent on the structure of the Weimarer Ausgabe, which bundles works together into single volumes that come from numerous time periods. There is also the problem of borderline works. For example, *Faust I* is labeled "late" because it was published after Schiller's death but we know its composition occupied much of Goethe's middle period (*Urfaust* is treated as its own work). Similarly, much

Table 2. List of quotation counts by genre in Goethe's corpus. The second column is normalized by the length of each genre.

Genre	Quotations	Quotations (per 100K words)
Prose	2847	182
Drama	1939	292
Critical	1747	261
Poetry	1682	520
Natural Science	983	158

Table 3. List of quotation counts by period in Goethe's corpus. The second column removes the critical and scientific writings, which are less reliable in terms of dates. The third column is normalized by the size of the period. Thus, given its small size, the early period is quoted at relatively higher rates than the other two periods.

Period	Quotations	Quotations (Reduced)	Quotations (Per 100K Words)
Late	5619	3668	162
Early	1907	1576	263
Middle	1584	1224	115

of the natural scientific writing wasn't published until the later collected editions though much of this writing dates to earlier periods.

And yet even if we remove the science writings and the critical writings on literature and the arts (which exhibit similar problems of dating), the late period still outnumbers the early and middle periods by a factor of 2:1. Once again, if we normalize by the size of the subcorpus (i.e., how many words there are to be quoted during each period) we see how the early period is now quoted much more frequently relative to the overall volume of writing. The middle or classical Goethe garners the least amount of attention no matter how you slice it.

The Language of the Unread

All of these measures can begin to give us a portrait of where a particular publication, in this case, the *Goethe Jahrbuch*, has lavished its quotational attention on a particular author. In this way, we can begin to understand the categorical orientation (or we might say less neutrally, bias) of the history of this publication. We see a strong orientalism evidenced by the *Divan*'s centrality, a strong orientation toward the prose works, particularly the autobiographical writings, and a lower than expected attention to the classical period. What these measures cannot tell us is what kind of language scholars are focusing on when they quote Goethe. What is the semantic portrait of

Fig. 2. Word cloud of the topics that are more likely to be used in nonquoted poetry.

Goethe that is being valued through quotation and what of Goethe's language is being left "un(re)read"?

To answer this question, we explore what we might refer to as the topic space of Goethe's quoted and unquoted poetry using a topic modeling algorithm.[11] Topic models allow us to estimate semantic constellations within a writer's work, akin to the study of what Ernst Robert Curtius called "historische Topik."[12] These tools are adept at identifying the likelihood of words co-occurring within poems and can identify larger-scale linguistic patterns. When doing so, we see that there are several topics that are more indicative of the unquoted poems, i.e., there are linguistic patterns that are latent within poems that are not quoted by Goethe scholars when compared to those poems that are quoted. Topics that tend to appear more often in the poetry quoted by scholars have to do with *nature, life*, and the *gods*, while the topics that are distinctive of the unquoted poems are related to themes of *love*, *girls,* the *day*, and the diminutive (*lieblich*, *Knabe*, *Jüngling* etc.) more generally (see Fig. 2). Indeed, if we look at the most distinctive individual words of each corpus (rather than "topics"), we see how words like *God*, *eternity*, *father*, and *world* are all significantly less likely to appear in nonquoted texts, while *heart*, *beauty*, and *liebchen* are much more likely to appear. These models suggest that there is indeed a strong gendered aspect to the history of Goethe quotation within the *Jahrbuch*, with a reliance on valuing Goethe's own language of universality and patriarchy over and above more domestic concerns that also find expression in his poetry.

Conclusion

We wish to conclude by drawing attention to both the limitations of what we have tried to do and also what we see as some potential future applications. The scholarly journal we have focused on is a small sample of a

much larger academic universe. The forms of attention we identified in the *Goethe Jahrbuch* are not necessarily indicative of wider trends in the field, though it would be interesting to understand the extent to which a journal like the *Goethe Jahrbuch* does or does not serve as a leading indicator of Goethe scholarship more generally. Nor would the larger scholarly context of quotations necessarily be representative of the sum total of the greater cultural practice of valuing Goethe (whether in Germany or a wider global context). Even if we expanded our view to a broader scholarly context, we would still only be capturing a very small subset of Goethe readers. And, of course, the practice of quotation is not a perfect mirror of reading. Indeed, it is in many ways an exceptional model—to reverse Price's formula from above—the passages we mark are the ones we truly want to remember and be remembered by. Quotation allows us to model reading and unreading in a particular way and to see the relationships between the language of the two as they play out in particular works. But, we are still left with all of the problems of understanding reading and its opposite, unreading, that we mentioned above.

Instead of making large-scale generalizations about the reception of Goethe's corpus or the meaning of the unread in his work, we see our work as a pilot project that aims to initiate a conversation about the practices of scholarly attention. The great unread is a useful heuristic for us because it captures the vast amount of Goethe's corpus (or any author's) that is ignored in critical discussions of the author. It highlights the differential and unequal nature of scholarly attention and encourages us to reflect on why this practice may or may not be valuable. When we continue to focalize a small subset of words from a single source, which is itself a small subset of a much larger universe, what values are we communicating to the reading public? By contrast, we think there is also value in recognizing all of the words and people who go unattended to and potentially unread. Call it the democratic turn in literary study. In drawing attention to our own attentiveness, we think computational methods can play a significant role in bringing about a change in whom and what we value.

McGill University

NOTES

1. Margaret Cohen, *The Sentimental Education of the Novel* (Princeton, NJ: Princeton UP, 1999) 23.

2. Franco Moretti, "The Slaughterhouse of Literature," *Modern Language Quarterly* 61, no. 1 (2000): 207–27.

3. Ted Underwood, *Distant Horizons: Digital Evidence and Literary Change* (Chicago: U of Chicago P, 2019).

4. Leah Price, "Reading: The State of the Discipline," *Book History* 7 (2004): 313.

5. Andrew Piper, "Sharing," in *Book Was There: Reading in Electronic Times* (Chicago: U of Chicago P, 2012) 83–108.

6. It is important to distinguish what we are doing from the practice of "citation analysis" because the study of citational networks between authors and works using computational methods represents an entire well-developed field of study (often referred to as bibliometrics). The study of "text reuse," on the other hand, for which we use the colloquial term "quotation," is valuable, especially for literary scholars, because it places the referential emphasis in the act of intertextuality on language rather than just the *work* from which language is drawn.

7. For other notable projects, see David A. Smith, Ryan Cordell, and Elizabeth Maddock Dillon, "Infectious Texts: Modeling Text Re-Use in Nineteenth-Century Newspapers," *Proceedings of the Workshop on Big Humanities* (2013); and Paul Vierthaler and Mees Gelein, "A BLAST-based, Language-agnostic Text Reuse Algorithm with a MARKUS Implementation and Sequence Alignment Optimized for Large Chinese Corpora," *Journal of Cultural Analytics* (March 18, 2019): DOI: 10.7910/DVN/2YYJ2B.

8. For the purposes of this project, we aggregate smaller poetic works into various genre groupings and larger works are broken down into volumes or sections based on the Weimarer Ausgabe of Goethe's works. They are then labeled by type of writing (poetry, drama, critical, scientific, and prose [both fiction and autobiographical]) and the period of Goethe's life when they were written (early, middle, late). In our representation of the corpus, there are a total of 254 documents. We consider this a preliminary model of Goethe's corpus that awaits fuller analysis and treatment (and work). All data and code are available at the following repository: https://doi.org/10.7910/DVN/EAXEET.

9. See https://github.com/JonathanReeve/text-matcher.

10. We can also burrow further into the data to see how different works are quoted differentially based on their parts. We already saw how *Faust I* and *II* are quoted with more or less equal frequency, which is interesting, but *Dichtung und Wahrheit,* for example, shows a skewed degree of attention, with considerably less attention going to book 1 (64) of than to the other books: book 2 (150), book 3 (150), and book 4 (116).

11. For a discussion of the place of topic models within the history of information, see Andrew Piper, "Topoi (Dispersion)," in *Enumerations: Data and Literary Study* (Chicago: U of Chicago P, 2018) 66–93. For the application of topic models to the study of the novel, see Matt Erlin, "Topic Modeling, Epistemology and the English and German Novel," *Journal of Cultural Analytics* (May 1, 2017): DOI: 10.22148/16.014.

12. E. R. Curtius, "Begriff einer historischen Topik," *Zeitschrift für romanische Philologie* 58 (1938): 129–42.