Words in a world of scaling-up: Epistemic normativity and text as data

Sayan Bhattacharyya

How does technology intersect with epistemology? How do the assumptions behind our analytical frameworks influence the inquiries they support and the results of those inquiries? In particular, while digital tools can provide useful means for the exploration and analysis of text as data, are there potential pitfalls that we need to be aware of when we use them for the exploration of less hegemonic languages? The present essay addresses these questions in the context of a digital humanities tool for discovering trends in large bodies of texts. When we consider power and coloniality in relation to text we usually think of how they shape the life-worlds that text depicts. However, as technology mediates our experience of interacting with texts in new (and previously unimaginable) ways, perhaps we need to start paying more attention to how the techno-social ensembles that constitute our experience of textuality structure our apprehension of texts that are themselves power-laden. For conceptualizing this understanding, Foucault's notion of an *apparatus* of knowledge is helpful: Foucault describes an apparatus as a system of relations between its constitutive elements, which together constitute a "heterogeneous ensemble consisting of, among other things, discourses, institutions…, regulatory decisions [and]…administrative measures," making up "the said as much as the unsaid" (194). To think in terms of knowledge-producing apparatuses as an analytical framework may help us to grasp the implicit assumptions underlying techno-social ensembles as well as to understand the structuring normativities that emerge from the interactions between their constituents.

Do apparatuses for producing knowledge render certain kinds of inscriptions invisible within their fields of discourse — inscriptions, especially, that have the kind of provenance associated with the rubric of postcoloniality? Madhava Prasad describes, with the help of a short story by A.K. Ramanujan, the trauma of erasure experienced by the postcolonial subject who finds herself in the West, when she discovers that elements from her cultural heritage have been rendered unrecognizable after being "processed" by the western academic machine (58) —a paradigmatic example of the kind of erasure or loss that is a recurrent theme in postcoloniality. Such losses typically tend to be experienced, as in the example above, through the *topos* of an outward migration of the postcolonial protagonist, or through the *topos* of a return — as, for example, in the instance of a novel widely considered a classic of postcolonial fiction, Chinua Achebe's *Things Fall Apart*. Whether or not the migration (or the return) is literal or allegorical, loss or erasure, in these instances, is a result of a change of epistemic framework, as protagonists move not only in space but also acquire new inhabitations that both support and demand a different regimen of normativity about knowledge production. This epistemic trauma as represented in literature arguably makes the postcolonial literature that expresses it, *world literature*, as the source of the loss is a translation or movement in space in the world — if we think of world literature, as David Damrosch argues, not as a particular canon of texts but rather as a mode of engagement with worlds beyond one's own time and space (281). In this essay, however, we address a different, but related, question. Rather than focus on "reading" individual works of literature — or even texts of any kind — in the usual, human, sense of the word "reading", our concern in this essay is with the possibility that, if we are not careful enough, the same kind of postcolonial erasure (if we call it so) may also occur when the apparatuses of knowledge production are computational tools that operate on archives consisting of large corpora of digitized text. Can this kind of apparatus, too, end up being complicit in the production of erasure or

loss when reading from the archive? We want to make two broad points pertaining to this complicity. Firstly, we want to point out how such a complicity can arise as computational tools for textual aggregation and analysis increasingly treat large textual corpora as big data and provide the means for extracting new knowledge from these large bodies of text. Our second point is that this same complicity also makes its appearance as world literature — in an extended sense of the word, as encompassing the world's texts — tends to become, in our time, a hegemonic, universal category.

This last point should help us in understanding that text considered as "big data" and literature considered as "world literature" share an epistemic commonality: they both have to do with the "worlding" of text in a way that is symptomatic of our current condition. The notion of "world" in the sense of an *ample*, all-encompassing extent — crops up as much in the phrase "world literature," as, arguably, it does in the phrase "big data" to denote a large corpus of text. But this notion of amplitude conveyed by phrases like "world literature" or "big data" is a compromised notion — on account of being far too premised on an epistemic framework both founded on, and favoring, the homogeneity of the implied categories they are based on. We[1] will show this by focusing on a tool developed recently for extracting and analyzing word usage across large corpora of digital text, that is, across text conceived as big data. While we focus on a specific tool here, the argument extends, *mutatis mutandis*, to many tools of this kind. We will see how, when large-scale repositories reconfigure collections of many individual texts into "big" textual data, then the combined actions of focusing, and of expansion of depth of field, that are afforded by large-scale inquiry can, in certain situations, render some non-western languages less visible. We wish to make it clear that this loss of visibility is not due to willful neglect or intentional suppression of any kind. As we will show, it is, instead, caused by a different problem, which is the following: hegemonic forms of representation of knowledge presuppose a homogeneous, rationalistic and standardized categorical order as their condition of possibility (— a presupposition that can get further amplified when small data is aggregated into the régime of big data). This has the consequence that an epistemic heterogeneity can be legible within a representational scheme encoded through standardized categories only when such subjectivity has already relinquished its heterogeneity to the normative assumptions of the dominant episteme, which privileges the homogeneous. We can think of this as an instantiation of the observation made by Gayatri Spivak in her essay, 'Can the Subaltern Speak?' (271); or perhaps more cogently, of a reformulation of that question by Jay Maggio, who shifts the focus of the question to whether the subaltern can be *heard* (419).

The computational tool that we consider in this paper is the HathiTrust Bookworm. Originally created at the Cultural Observatory at Harvard University, Bookworm is an interactive tool developed by Erez Lieberman Aiden and Benjamin Schmidt for the visualization of trends within repositories of digitized texts by means of "culturomic" analysis (Michel *et al.* 176). These trends do not necessarily have to be chronological, but can be trends across *any* category (not just time) that can be expressed by means of a metric. The philosophy behind the Bookworm tool is to abstract across categories and express the variation of some attribute of the corpus across those categories as a plot. The HathiTrust Bookworm, the particular tool that interests us in this essay, is a specific adaptation and instantiation of Bookworm for visualizing content from the HathiTrust Digital Library, which contains almost 15 million volumes of digitized text (Auvil *et al.* 2015). Search engines are adept at finding individual texts within a digital library, but the HathiTrust Bookworm performs a quite different task — it *abstracts* across categories within a digital library. The generation of these abstractions can be considered a form of "distant reading," a kind of reading in which distance from the text is the condition of knowledge, and the act of reading is performed on units that can be much larger than the

individual text, such as an entire corpus of texts or subsets of it. This allows for the discovery of large-scale patterns within the unit through distant reading (Moretti 48).

Generation of such abstractions by the current instantiation of the HathiTrust Bookworm takes place in the context of a word that the user specifies to it. The generation of the abstraction is carried out through the variation of some attribute of that word across the categories that defined by any chosen categorization scheme. For example, for the HathiTrust Digitial Library, a canonical categorization scheme is the organization of the digital library by Library of Congress classes as metadata. The particular attribute of the occurrence of a word across categories that the HathiTrust Bookworm chooses for visualization is its normalized frequency of occurrence within books in HathiTrust across those categories. HathiTrust+Bookworm can work both with discrete sets of categories, such as Library of Congress classes, as well as with continuous categories, such as time. The plots can show the relationship between two variables, as in a two-dimensional time-series, or the relationship between three variables, as for example in a heat map. (A relationship between more than three variables is difficult in practice (although possible in principle) to display in the two dimensions available on the printed page or on the screen.) A two-dimensional time-series plot of the normalized frequency of occurrence of a word over a chronological range provides a sense of how the relative occurrence of that word or phrase has varied over the specified time range. The user can generate visualizations consisting of plots that show how change in social context correlates with the change in preponderance of one word over another within the corpus. For example, Fig. 1 shows how the normalized numbers of occurrences[2] of two words have diverged over time. Fig. 2 shows a relationship between three variables by means of a "heat map." The color (so-called "heat") "intensity" in each cell in the heat map represents the normalized number of occurrences in books of the queried-for word corresponding to that cell in the map. Each cell in the map is constituted by a Library of Congress class along the y-axis, and by another category, book size, along the x-axis. The latter is, in this example, a discrete (discontinuous) category— the number of pages in the book, "binned" into the classes 'small', 'medium', 'large', *et cetera.*

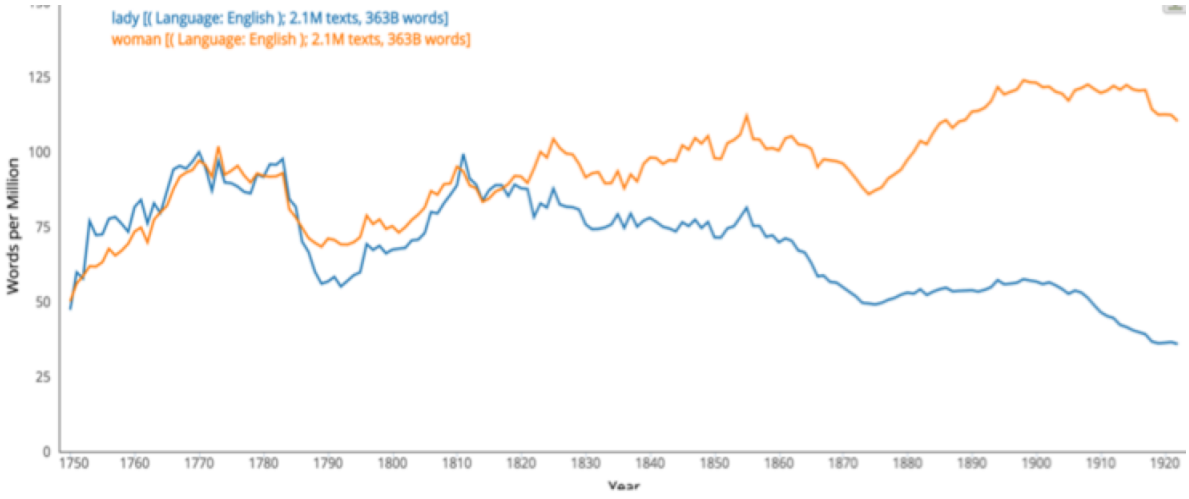

Fig. 1

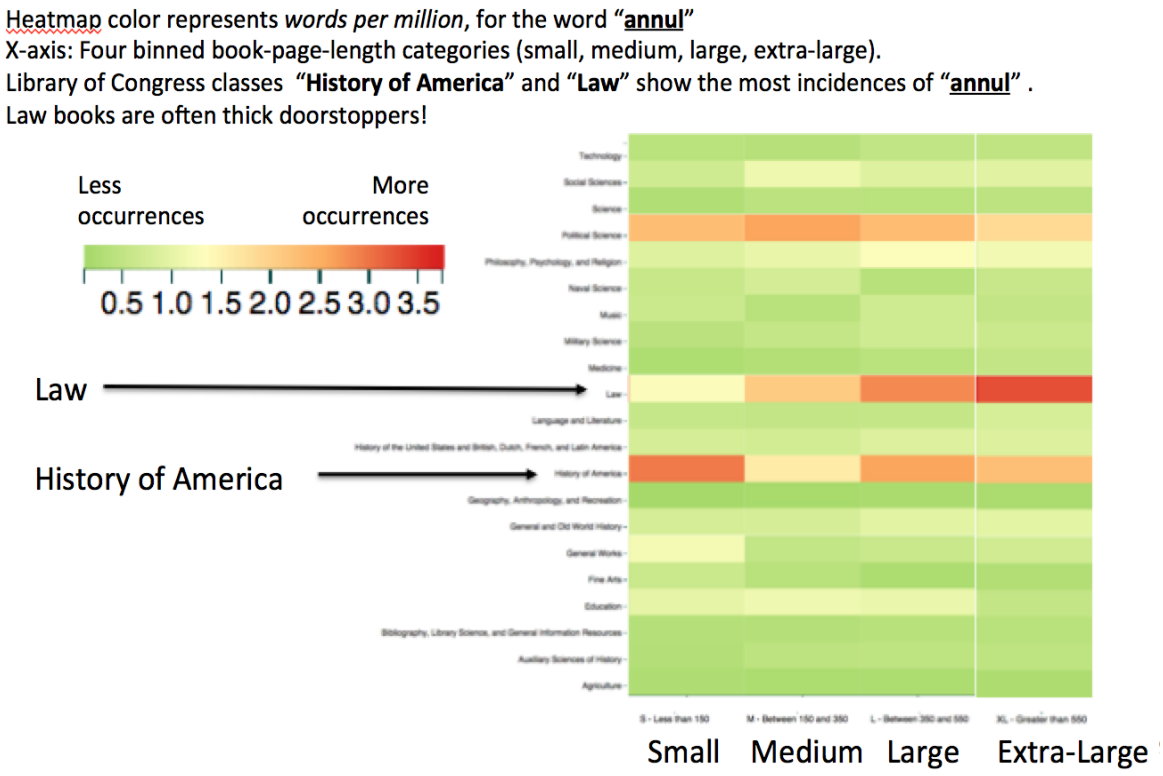

Fig. 2

While the HathiTrust Bookworm provides an entry point for useful explorations of the HathiTrust collection of digitized books, there are limits to how useful such a tool can be for exploration of words from less hegemonic languages. The limit that we are interested in here does not simply concern the relative paucity, within the world's library collections, of books written in the world's less hegemonic languages (although this is of course an important, but separate, issue in and of itself) or the fact that the quality of optical character recognition for digitized text in languages written in non-roman scripts is typically lower than that for languages in roman scripts. These, of course, are important and interesting problems in themselves, but what we are interested in here is, precisely, the complex intersection of technology with epistemology: how tools, (and, by extension, the analytical frameworks upon which the tools are based) shape and inflect the kinds of investigations that can be carried out with the help of those tools (and frameworks). We draw attention to an interesting problem that arises when the queried word in a tool like the HathiTrust Bookworm is one that is from a non-European language, but which occurs within European-language texts — with the word occurring in the text in the roman alphabet, that is, in *transliterated* form. We found that the occurrence of such words was being underreported, and sometimes not being reported at all — an observation that was initially unexpected to us (see Fig. 3, where we illustrate this with a screenshot[3]). Investigating the reason behind this leads to understanding that a rather surprising interplay involving historical and contemporary practices concerning representation is producing this state of affairs.

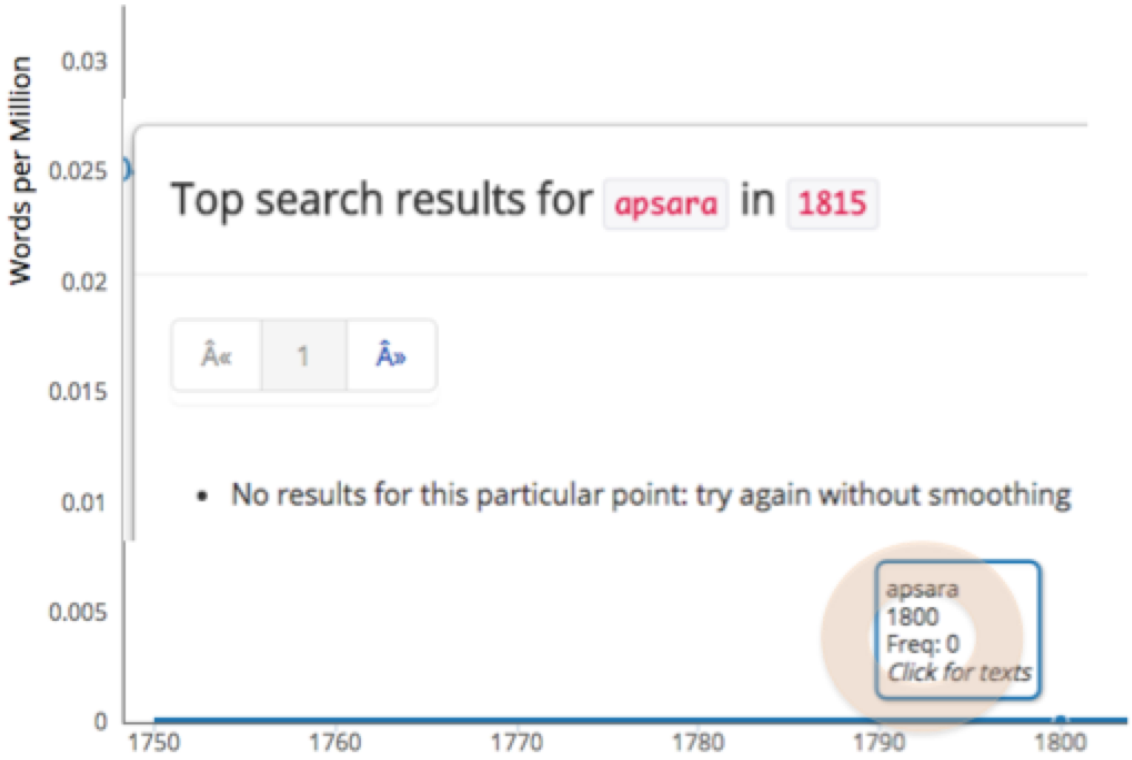

Fig. 3

The problem is caused by the fact that a tool like the HathiTrust Bookworm has to depend upon an index for looking up the queried word and returning its frequency of occurrence, which means that, ideally, the index needs to have an entry for every single unique word in the corpus that could potentially be displayed through the tool. However, words that are too infrequent in the corpus are usually not stored in such an index as the one that the HathiTrust Bookworm maintains for itself, as indexing so many words would cause performance problems, slowing down retrieval in real-time for plotting purposes.[4] So, low-frequency words cannot be put in the index, and have to be treated, effectively, as if they *never* occurred in the corpus. This has the result that, as the size of the collection is scaled up, low-frequency words are disproportionately affected. Now, one could argue, of course, that the decreasing cost of storage and improvements in computing power (speed of computation and retrieval) that are constantly occurring will render this scaling-up problem moot in the future — as, after all, even assuming that the text collection grows exponentially, the number of *distinct* words in the text collection is likely to increase at a much slower rate than the exponential rate of growth of computing power. There is, however, a different and more complex problem, that is of interest to us here.

This problem has to do with variant spellings of words. This is a problem that is particularly salient in the case of words from those languages that *are* ordinarily written in their own, non-European, scripts (in which, of course, the words have more or less standardized orthography). Such words, when occurring in books written in European languages, would typically be in roman script, that is, they would be transliterated into roman script when they are mentioned in a European-language book. This is typically the case for words from South Asian languages — these languages are written in various (non-roman) alphabets, but when a word from such a language happens to be

mentioned in a European-language book, that word would usually occur, not surprisingly, in roman-script transliteration, rather than in its own script — as, obviously, typesetting cannot accommodate these multiple scripts. It is instructive to contrast this situation with the case of non-western languages that are written in European (usually roman) scripts, such as, for example, Bahasa Indonesia, or Turkish — in the case of the latter, since the time that Turkish started to be written, following Kemal Ataturk's reforms in the 1920s, in the roman script instead of in the *nastaʿlīq* script previously used in the Ottoman empire. In the case of these latter languages, such as Bahasa Indonesia or Turkish, there is usually a single unique roman-script spelling of a given word, and this word usually also has a standardized spelling.[5] However, in the case of South Asian languages, which have been written in non-European scripts for a long time before the European conquest, a single, standardized way to transliterate a word into roman script usually has not been in existence, especially in the earlier years of the experience of colonial encounter. Thus, the same word in a South Asian language could typically be transliterated into roman-alphabet script in multiple, different ways. (The number of possible variations would be greater, of course, the more phonemes there
happen to be in the word.)

This difference has consequences. Suppose that a tool like the HathiTrust Bookworm needed to support the retrieval of frequencies of occurrences of South Asian language words that occur transliterated in roman script within the HathiTrust collection. The need for supporting such a specialized function may arise, for example, if the tool is to be used in a class that specifically focuses on South Asia; in fact, we discovered this issue in the first place when we used the HathiTrust Bookworm in a class.[6] The intuitive way to provide such support would be to make a special dispensation — by accommodating, in the index used by the tool, the words belonging to the special class, including them *in spite of* their falling below the threshold of frequency of occurrence ordinarily used as the lower bound for inclusion in the index in the standard case. This will, indeed, lead to the recuperation of these words from invisibility. However — and this is the crucial point here —this is going to work well only in the case of those non-European languages that are written in roman script to begin with. Let us see why this is so. Consider, for example, Bahasa Indonesia. One could create an entry in the index for each word in Bahasa Indonesia that occurs, say, in a dictionary of Bahasa Indonesia, or one could create an entry in the index for each unique word from a word list consisting of each word encountered in some substantial corpus of Bahasa Indonesia texts. The point is that there is a way to create a reasonable list of Bahasa Indonesia words, as the notion of a roman-script Bahasa Indonesia word is fixed and stable. Let us take the example of a random word from Bahasa Indonesia: the word denoting the number "ten" in Bahasa Indonesia is the word "sepuluh." A search for the word "sepuluh" in the full text of all of the HathiTrust digital library's books published prior to 1923 showed us 206 occurrences of this word. It is true that 206 is a small number — and, in fact, in an earlier iteration of the index for the HathiTrust Bookworm, this number had proved too small for the word to be included in the index. However, Bahasa Indonesia being a language written in roman script as a matter of course, the word "sepuluh" would not typically show orthographic variation when it occurs in text. So the word, whenever it occurs, is much more likely to appear with the fixed, standard spelling rather than appear as other orthographic variants. And even if such a word *were* to show significant fragmentation among its occurrences due to the existence of distinct orthographic variants (as could happen, for example, for occurrences in texts from the chronological period before the orthography of the language was standardized) it would still be relatively straightforward to include all the variants in the index under the same head— for example, by collecting the different variations in spelling of the word from a dictionary. That Bahasa Indonesia is written in *roman* is what makes the crucial difference here — a similar dictionary that records different existing variations of a word in *roman*, transliterated spelling would not exist for languages that are not ordinarily written

in roman script. In short, a word like "sepuluh" in a language like Bahasa Indonesia that is ordinarily written in roman script can, in principle be recuperated from invisibility by simply forcing its inclusion in the index (easily collecting and including any variations in spelling under the same head, if needed), leading to the word being reported, and thus made visible, by the HathiTrust Bookworm tool when the tool is used to look for occurrences of the word within European-language texts.

However, the situation is very different when the tool is used to plot, within books in European languages written in roman script, occurrences of words transliterated from those languages that are usually *not* written in hegemonic scripts such as roman. Such words can, as we mentioned earlier, typically be transliterated into roman script in multiple, different ways. In the late eighteenth to the early twentieth centuries, in particular, there usually were no schemes in existence for mapping such transliterations in a standardized way.[7] The key issue here is that it is not possible, in general, to anticipate, *a priori,* all such possible variations. Had it been possible, one could have aggregated all such variants so as to programmatically spawn separate queries on the index for each variant (to obtain the number of occurrences for each variant) and then lumped the results together. The heterogeneity expressed in variant orthographic transliteration, which prevents us from doing so, is a kind of fluidity, impossible to accommodate within brittle categories that are determinate and discrete. However, normative categorization schema, whose genealogy, in the modern world, can be traced back at least to the European Enlightenment and possibly even further back, are, indeed, precisely such determinate, discrete and brittle categories. The only way that the substantial content of this heterogeneity can become legible is if it has always already been coded/translated (or, as in this case, transliterated) in terms of the hegemonic forms of knowledge-organization embodied in the apparatuses of knowledge production and storage constituted by the library catalog (and its present-day descendent, digital data). This is an instance, within what we nowadays call digital humanities, of how content that is of a subaltern status must always already conform to the epistemic assumption that underlies hegemonic régimes of knowledge organization before such content can be legible.

While we can broadly think of this phenomenon as an instantiation, in the new era of "big data", of Spivak's key insight concerning the epistemic legibility of the subaltern, Yves Citton's work relating epistemology, information and attention together can provide a possible theoretical lens to sharpen our understanding of the process. Citton borrows McKenzie Wark's notion of vectoralism — control over the vectors along which information is abstracted, by way of analogy with the corresponding ways in which value was abstracted in pastoralism through control over the pastors in agrarian society and in capitalism through control over capital in commodity production (Wark, paragraph 025) — and extends the notion of the vectorial from power over information alone to, more broadly, power over attention and hence on visibility and knowability, as well: vectoralism, formulated in this way, is the realization of value in countable terms (Citton 78). Conversely, then, only that which can be easily counted can have value in a régime of vectorialization. This notion of countability, followed by attention to matter only in its digitized form, as underlying value in a world of digitization, is premised on the digital as the latest formulation of what Bernard Stiegler describes as grammatization — the historical phenomenon that transforms the analog and heterogeneous sensory continuum to discrete, digitized bits, by imposing a grammar, in the sense of a standardization, on that continuum (8).

Kitchin and Lauriault have noted that the production of academic knowledge has progressed over the past few centuries using what may be called "small data" studies — based on very selectively sampled data generated to answer specific questions (463). In the humanities, this has typically

consisted of close reading of small bodies of text. However, the feasibility of using computational techniques based on big data is increasingly enabling the analysis of large bodies of text at scale. This has had the consequence that "small data" are now often being aggregated into big data through the development of new data infrastructures that, as Kitchin and Lauriault put it, "pool, scale and link" small data into larger datasets, opening them up to large-scale analysis (463). What is at play here, then, is the interaction between, on the one hand, apparatuses of knowledge — printed books bearing the trace of relations of power and colonization within their contents — and, on the other hand, the aggregation and connectivization of extensive "small data" elements into "big" data in the form of individual digitized books from many libraries being combined into a single text archive that can be algorithmically processed. Such a text archive constitutes a contemporary apparatus of knowledge: a data infrastructure. The illegibility/invisibility that we have described is caused by the interaction between two knowledge apparatuses: the historical archive elements created under conditions of (post)coloniality, and the digital, algorithmic tool that retrieves and visualizes results computed from the archive elements in response to queries. As Rebecca Walkowitz suggests, the comparative method in literature should focus not only on text but also on materiality of the text (568) — an idea expanded upon by Jessica Pressman to include not only materiality but also text understood and analyzed in terms of *process* ("Electronic Literature as Comparative Literature" 2014). Attempting to understand how the underlying process works in a digital humanities tool for text analysis, such as the HathiTrust Bookworm, and how these processes inflect the results of the inquiries we perform with them, is a step in that direction.

It may also be worth pointing out that the notion of world literature itself raises in its own way questions similar to those relating to the ones we have discussed in this essay: the difficulty of representing sensory continuity through discrete, determinate objects (and the loss that our descriptions of the world suffer when we are forced to choose a single, discrete and standardized representation in the interest of grammatization (to use Stiegler's term), and which in turn serves the interest of vectorialism through the efficient apportioning of attention in countable form. Even though this issue is, of course, different from that raised by algorithmic tools for text, similar questions related to epistemic legibility and power relations are at play in both. Thus, the issues that we discussed in the context of a digitized tool for text analysis may provide a useful analogy to think about text in a global context. Chigozie Obioma has drawn our attention to the dilemma facing Nigerian writers (and, by extension, any writers writing from the global periphery) writing in a "global" language like English in making a choice as to whether to describe objects and concepts unfamiliar to a global readership by the use of a single, unglossed, indigenous-language word, which may fail to make its referent legible to the western reader who lacks the context to understand the various associations of the word. He provides the example of *molue*, the name of a type of bus from Lagos, Nigeria, whose description, he writes, "is not complete without a description of its unique sound" (Obioma). Choosing a unique, determinate word-equivalent, namely the transliteration "molue" in roman, fails to do justice to the sensory continuum embodying the knowledge that this word-concept encompasses. Writers from the global periphery who write in English but about non-anglophone cultures often deal with this issue in a creative way, as Namwali Serpell points out, by using different kinds of glosses. When using such a non-English word from the cultural context being written about, a writer in this situation would often choose to use multi-word glosses rather than a single-word English equivalent for the non-English word (Serpell). Attempting to capture the sense of such a non-English word in a single, determinate, English-language-equivalent word would cause a loss of meaning. Chimamanda Adichie notes that not only writers from marginal parts of the world, such as her native country, Nigeria, but also writers from marginal groups in general are, however, often told that they have to give the reader "some form of entry." This usually means, she writes,

"toning down" language, by selecting words that would be immediately familiar to their readers (Adichie).

The examples constitute instances of the kind of standardization or grammatization enacted by apparatuses of knowledge — constituted, in this case, by the political economy of the publishing industry and the book trade, and the unintended effect that it has in making subaltern voice inaudible unless already translated (in this case, literally) into a register that conforms to epistemic normativities. Derek Attridge's suggestion that we should read without imposing a fixed set of norms on the text and that we should always inform rule-governed reading with an unconditional hospitality and openness (Attridge 305) can provide a possible pointer as to how, in our exploration of digital text corpora, we should always be careful not to make our practices become unintended prisoners of epistemically determined normative assumptions.

The conceptual issues that we have considered in this essay — even though they arise out of a specific situation in the context of a particular digital humanities tool — are, thus, homological to these larger issues relating to text, language, and questions of power and subalternity in a world context.[8] It is worth re-emphasizing that this does not in any way constitute a criticism of the tool itself. The HathiTrust Bookworm is a highly useful tool and it has been of great utility in our classes, and others have found excellent use for it in their research. The issues that we have discussed do not detract from the usefulness of the tool itself, and they do not constitute an indictment of digital humanities tools in general, either. What we have tried to show in this essay is that the novelty of such tools, however, provides a renewed opportunity for reflection on older, recurring questions.

Acknowledgements: I gratefully thank Christi Merrill at the University of Michigan, Ann Arbor, Jim English at the University of Pennsylvania, and the anonymous reviewer of the paper, for their generosity in providing useful suggestions and comments on earlier drafts of the paper. I also owe a debt of gratitude to the members of the HathiTrust+Bookworm team, especially Peter Organisciak and Loretta Auvil at the University of Illinois, Urbana-Champaign, and Ben Schmidt of Northeastern University, for their conceptual and technical work in creating and developing the HathiTrust Bookworm tool. Erez Lieberman Aiden and Ben Schmidt, were the original creators of Bookworm. I am especially grateful to Christi Merrill for the opportunity to use the HathiTrust Bookworm tool in the undergraduate classes taught by her in Asian Languages and Culture and in Comparative Literature at the University of Michigan; it was in course of that activity that Christi first discovered the seeming anomaly when searching for the word *apsara*, which occurs in several South Asian languages, among English-language texts — which led to the investigation that eventually led to this paper. Peter Organisciak and Loretta Auvil provided very helpful information about the technical matters underlying the tool. Any mistakes are, of course, my responsibility entirely.

Notes:

[1] In this essay, we use the first person plural ("we") throughout, instead of the first person singular ("I"). This is not merely a rhetorical convention of the authorial "we"; there is a more substantive reason for the choice. Work related to the digital humanities happens as part of team efforts, in collaboration with others; and the use of the plural seems more appropriate for that reason.

[2] The normalized number for a year is the number of occurrences in a year, per total number of books in the HathiTrust corpus that were published in that year.

[3] The screenshot in Fig. 3 is of the graphical user interface of the HathiTrust Bookworm tool from early 2016. It shows what the tool reported when we used it to query the word "apsara", a Sanskrit word that occurs in various other South Asian languages as well What we observe in Fig. 3 is that the tool is reporting the normalized frequency of occurrence of the word as zero (the flat blue line hugging the x-axis) — a case of under-reporting. As we see from the figure, clicking on the plot at particular years also reported no texts in which the word *apsara* had occurred. (That under-reporting was taking place was confirmed by searching the HathiTrust text corpus using a different procedure to confirm that the word did occur in at least some texts over the time range of interest.)

[4] This is because word occurrence tends to follow a power-law distribution in all languages— that is, a short "head" of the distribution consisting of just a few words with a high frequency of occurrence, followed by a very long "tail" of the distribution consisting of many words with a low frequency of occurrence (Piantadosi 1112).

[5] Here we are not considering issues such as stemming or lemmatization, that is, identifying the stem, or base form, of *derivationally* related words (derived from the base form). The question of lemmatization is an issue orthogonal to, and separate from, this discussion, and does not affect our argument.

[6] We used the HathiTrust Bookworm in a class taught by Christi Merrill at the University of Michigan, Ann Arbor, in 2016. We have described elsewhere how we used the HathiTrust Bookworm tool in the class ("Big-Data Oriented" 1).

[7] As a case in point, consider the last name of the author of this essay itself: the spelling *Bhattacharyya* is not the only possible transliteration, and this name, which is a Sanskrit compound word, has been, and still is, spelled in English in various possible ways: *Bhattacharjee*, *Bhattacharya*, *Bhattacharjya*, *Bhattacharja*, and several other variants.

[8] This is not to deny that reading literature is, of course, different in a fundamental way from analyzing text — although, as I have argued elsewhere, it is instructive to look for similarities between the usual practice of reading and its technologically mediated posthuman variants ("A Fragmentizing Interface" 62).

Works Cited

Adichie, Chimamanda Ngozi. "PEN World Voices Festival of International Literature, New York City". *YouTube*, May 3, 2017, http://youtu.be/yiX5XvykVSk. Web. Accessed August 31, 2017.

Attridge, Derek. *The Work of Literature,* Oxford: Oxford University Press, 2015. Print.

Auvil, Loretta, Erez Lieberman Aiden, J. Stephen Downie, Benjamin Schmidt, Sayan Bhattacharyya, and Peter Organisciak. "Exploration of Billions of Words of the HathiTrust Corpus with Bookworm: HathiTrust + Bookworm Project", *Digital Humanities 2015 Conference* (*DH 2015*), Sydney, Australia, 29 June - 3 July 2015. goo.gl/W11teK. Web. Accessed Jun. 10, 2017.

Bhattacharyya, Sayan, Christi Merrill, Peter Organisciak, Benjamin Schmidt, Loretta Auvil, Erez Lieberman Aiden, and J. Stephen Downie. "Big-Data Oriented Text Analysis For The Humanities: Pedagogical Use Of The HathiTrust+Bookworm Tool", *Digital Humanities 2017 Conference (DH 2017)*, Montreal, Canada, Aug. 8-11, 2017. https://dh2017.adho.org/abstracts/414/414.pdf. Web. Accessed Aug. 29, 2017.

Bhattacharyya, Sayan, Peter Organisciak, and J. Stephen Downie. "A Fragmentizing Interface to a Large Corpus of Digitized Text: (Post)humanism and Non-consumptive Reading via Features", *Interdisciplinary Science Reviews* 40 (1), (special issue on 'The Future of Reading'), Mar. 2015, pp. 61-77. http://www.tandfonline.com/doi/full/10.1179/0308018814Z.000000000105. Web. Accessed Aug. 31, 2017.

Chaudhuri, Amit. "On Literary Activism", opening remarks at 'The Making of the Literary' symposium, University of East Anglia, India Workshop, Jadavpur University, Kolkata, India, Dec. 2, 2014. http://ueaindiacreativewritingworkshop.com/symposium-on-literary-activism/. Web. Accessed Aug. 16, 2017.

Citton, Yves. *The Ecology of Attention*, Cambridge, England: Polity Press. Print.

Damrosch, David. *What Is World Literature?*, Princeton: Princeton University Press, 2003. Print.

Foucault, Michel. "The Confession of the Flesh", interview, in *Power/Knowledge: Selected Interviews and Other Writings,* translated by Colin Gordon, Leo Marshall, John Mepham and Kate Soper, ed. Colin Gordon, Pantheon Books, 1980, pp. 194-228. Print.

Kitchin, Rob and Tracey P. Lauriault. "Small Data, Data Infrastructures and Big Data", *GeoJournal,* 80 (4), 2014, pp. 463-475. Print.

Maggio, Jay. "Can the Subaltern Be Heard?: Political Theory, Translation, Representation, and Gayatri Chakravorty Spivak", *Alternatives: Global, Local, Political,* 32 (4), Oct.-Dec. 2007, pp. 419-443. Print.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science*, 331 (6014), Dec. 16, 2010, pp. 176–82. doi:10.1126/science.1199644. Web. Accessed Jun. 11, 2017.

Moretti, Franco. *Distant Reading*, London: Verso, 2013. Print.

Obioma, Chigozie. 'Who Should I Write for – Nigerians, Africans, or Everyone?' *The Guardian*, October 14, 2016. https://www.theguardian.com/books/2016/oct/14/how-african-writers-can-bring-local-language-to-life-for-all. Web. Accessed August 25, 2017.

Piantadosi, Steven T. "Zipf's Word Frequency Law in Natural Language: A Critical Review and Future Directions." *Psychonomic Bulletin & Review* 21 (5), Oct. 2014, pp. 1112–1130. https://link.springer.com/article/10.3758/s13423-014-0585-6. Web. Accessed August 31, 2017.

Prasad, Madhava. "On the Question of a Theory of (Third World) Literature", *Social Text* 31/32 (1992), pp. 57-83. Print.

Pressman, Jessica. "Electronic Literature as Comparative Literature: The 2014-2015 Report on the State of the Discipline of Comparative Literature", American Comparative Literature Association, June 14, 2014. https://stateofthediscipline.acla.org/entry/electronic-literature-comparative-literature-0. Web. Accessed Aug. 26, 2017.

Serpell, Namwali. "Glossing Africa", *The New York Review of Books NYR Daily*, August 21, 2017. http://www.nybooks.com/daily/2017/08/21/glossing-africa. Web. Accessed Aug 28, 2017.

Spivak, Gayatri. "Can the Subaltern Speak?", *Marxism and the Interpretation of Culture*, edited by Cary Nelson and Lawrence Grossberg, Urbana: University of Illinois Press, 1988, pp. 271-313. Print.

Stiegler, Bernard. *For a New Critique of Political Economy*, trans. Daniel Ross, Cambridge, England: Polity Press, 2010. Print.

Walkowitz, Rebecca. "Close Reading in the Age of Global Writing", *Modern Language Quarterly* 74 (2013), pp. 171-195. Print.

Wark, McKenzie. *A Hacker Manifesto*, Cambridge, Mass.: Harvard University Press, 2004. Print.

Sayan Bhattacharyya
Postdoctoral Fellow
University of Pennsylvania
sayanb@sas.upenn.edu