CA Journal of Cultural Analytics

July 17, 2019

On the perceived complexity of literature. A response to Nan Z. Da

Fotis Jannidis[a]

[a] *Universität Würzburg, Germany*

ARTICLE INFO

Article DOI: 10.22148/001c.11829

Journal ISSN: 2371-4549

ABSTRACT

At the center of Nan Z. Da's article is the claim that quantitative methods cannot produce any useful insights with respect to literary texts: "CLS's methodology and premises are similar to those used in professional sectors (if more primitive), but they are missing economic or mathematical justification for their drastic reduction of literary, literary-historical, and linguistic complexity. In these other sectors where we are truly dealing with large data sets, the purposeful reduction of features like nuance, lexical variance, and grammatical complexity is desirable (for that industry's standards and goals). In literary studies, there is no rationale for such reductionism; in fact, the discipline is about reducing reductionism."

At the center of Nan Z. Da's article is the claim that quantitative methods cannot produce any useful insights with respect to literary texts:

> CLS's methodology and premises are similar to those used in professional sectors (if more primitive), but they are missing economic or mathematical justification for their drastic reduction of literary, literary-historical, and linguistic complexity. In these other sectors where we are truly dealing with large data sets, the purposeful reduction of features like nuance, lexical variance, and grammatical complexity is desirable (for that industry's standards and goals). In literary studies, there is no rationale for such reductionism; in fact, the discipline is about reducing reductionism. (638)

From this quote, one could assume that the article is concentrating on showing the "drastic reduction of literary, literary-historical, and linguistic complexity" at work in the more than 15 studies she mentions. But her criticism, especially in the eight replication studies, is much more traditional—traditional in the field of data science. For example, she points out that Ted Underwood should have done B instead of A, because A "does nothing for his objective" (608). Ted Underwood explained that he actually did B, which Da conceded in her answer (Da 2019b).

In other cases she points to problems due to different corpus sizes or the instability of results due to the usage of different stopword lists. All this is good and has to be done—and errors can and will happen on all sides. But this is not the crucial point. Even if she could have shown that all CLS studies she analyzed are in some way flawed according to the standards of inference statistics - and to emphasize this: they are not - there is no way which leads from that finding to her very sweeping statements: "The nature of my critique is very simple: the papers I study divide into no-result papers—those that haven't statistically shown us anything—and papers that do produce results but that are wrong." (605) Her analyses of the studies don't support her claim because the quality of specific studies cannot be the basis of an argument for the question of whether quantitative methods can be applied to literary texts with valid results.

The bonmot at the end of the quote cited above, "the discipline is about reducing reductionism", receives its plausibility from the confrontation with the professional sectors: industry dumbs down language for its purposes, while literary scholarship does the opposite. But this opposition immediately looses plausibility when one realizes that these methods are also used in other fields of research for the purpose of research into complex matters—for example computational linguistics, sociology, psychology, bio-informatics and more. Most research fields would also claim that one of their goals is not to 'reduce complexity'.

The rhetoric of Da's article is based on obscuring the boundaries between the very general statements about CLS on the one side and the very limited reach of her methodological criticism on the other. Basically she says over and over again: all practitioners in the field of computational literary studies are incompetent—and analyzing literary texts with computational means wouldn't work anyway. The fundamental problem of her essay is that she tries to combine these two quite different statements into one argument, where the errors are supposed to prove the point about the futility of the approach. At the beginning of section 6, for example, she states: "CLS has no ability to capture literature's complexity." (634) But in the sentences following this general statement, she delves into one specific paper by Mark Algee-Hewitt, who uses entropy of bigrams as a measure for literary complexity and finds a correlation with canonicity. Her criticism advances in two steps: she explains what entropy measures and emphasizes the difference between higher diversity of words and more complex meaning. But as

Algee-Hewitt didn't claim to capture "meaning," she concedes the point only to explain how he—in her view—miscalculated the differences. This second step maybe fruitful inside of the field CLS, but doesn't show that "CLS has no ability to capture literature's complexity", because there is no logical way to move from an error in a calculation of a researcher to a general statement about the fruitfulness of a research field in general. (This also applies to Da: The fact that a calculation of hers includes an error—as some have mentioned—, doesn't say anything about the validity of her arguments in general.). Her analysis of statistical errors does nothing for her objective, *if* this objective is the argument that CLS is a field of research that cannot use quantitative methods because of the specific quality of its object, literary texts.

If you look at the two sides of her argument, the errors of CLS and the futility of its research program, it is interesting to see that most of the text is about the errors, but there are only a few arguments bolstering the second claim. In the following, I concentrate on this second claim. As far as I can see, there is only one fundamental argument she employs, which we saw quoted above: complexity. And then there is a second group of methodological arguments, which are derived from the papers she analyzes, but which she at least in parts transforms into general objections to the application of quantitative methods to literary texts: these center around questions of operationalization, the high dimensionality of language data, and 'proper' inference statistics.

Complexity

Nan Z. Da doesn't explain why literature is supposed to be especially complex, and maybe she doesn't have to because it would be fair to assume that this is a shared belief in literary studies. However, to make sense in this context it is not enough that literature is very complex. It has to be *singularly* complex, given that many other research fields with complex objects use quantitative methods. Sociology tries to describe whole societies, including artists and artworks, just as psychology tries to understand the psyche of individuals as well as groups, including those who produce and perceive works of art. So, yes, a literary text is probably infinitely complex, and the society in which this text is produced, distributed and read, is also infinitely complex, as is the human psyche. But these infinities probably don't have the same size, just like natural and real numbers. In light of these comparisons, it is very hard to believe an argument that literature is *singularly* complex, that *only* in the case of literature there can be no quantitative research.

But this is the underlying assumption which permeates most of the other arguments. When Da is talking about the use of NLP methods her argument basically goes like this: 1) CLS is not using better and more complex methods than those used in Computational Linguistics. 2) These methods don't work for literary texts, because they only work with sets of simple and similar texts. 3) Even if they work on the same level as non-literary texts, this is not enough for literature. "You quickly run into a data scarcity and data complexity problem with literature." (636) The first point is probably an effect of her sample of CLS texts. I will come back to that at the end. The second point is actually an empirical question. Da states: "Speech tagging is extremely inaccurate for literary texts." (635) Tellingly enough in this reference-rich text there is no reference backing up this claim. Does the complexity of literary texts really reveal itself in something like part-of-speech tagging? Yes, she says: "Lexical, syntactic, and grammatical ambiguities make it difficult for an algorithm to know whether a word is a participle or a gerund, if an adjective is a noun, or if entire phrases are functioning as a single part of speech." (635) Maybe this is true, if you look at modern poetry or experimental prose, but not very likely if you look at most of the fictional prose. But again, it is an empirical question, and I can only offer one piece of empirical evidence here: In our studies on character references in German literary prose, in which we annotated texts from the 19th and the beginning of the 20th Century, we didn't see especially high error rates after we finished the tedious business of creating training corpora for our domain. What we did find were different distributions compared to non-fictional prose (Krug at al. 2018). So even if the material is definitely different, at least the one NLP method we used worked fine. But a key point here is that these assertions can and ought to be tested.

The third point seems to be a variant of the complexity of literature argument: "Tagging errors and imprecision in NLP do not sufficiently degrade the extraction of information in many other contexts, but they do for literature." (636) The concrete form of the argument is rather unclear, because all of her observations up to this point support argument 2, that NLP tools cannot work as precisely on literary texts as on non-literary texts. But they do not explain why a working NLP tool with high, but not perfect, reliability wouldn't suffice for literary texts. Why

95% precision with high inter-annotator agreement is good enough for linguistics but not for literary studies remains unclear.

This lack of coherence between her arguments is quite typical for the whole text. The first argument attacks CLS because they don't use more sophisticated methods, methods which she claims in her second argument, do not work for literary texts anyway. And the same lack of logical connection can be found for argument two and three. She is just stacking them up like ramparts around the center of a castle to make sure that it is impossible to reach it.

Statistics

As mentioned above Da repeatedly criticizes the foundation of CLS studies:

> No matter how fancy the statistical transformations, CLS papers make arguments based on the number of times x word or gram appears. (605)
>
> Therefore all the things that appear in CLS—network analysis, digital mapping, linear and nonlinear regressions, topicmodeling, topology, entropy—are just fancier ways of talking about word frequency changes. (607)
>
> No one has ever said, though, that consistent word frequency is what distinguishes Shakespeare's comedies from tragedies, tragedies from histories, and so on—and no one would ever say that because such distinctions cannot be captured with word frequencies. (622)

Is she right? Yes, a lot of work in CLS and also in computational linguistics uses token frequencies, where a token can be anything from characters to words, taken either as 1-grams, 2-grams, or 3-grams, etc. Is this bad? Da contrasts it explicitly and implicitly with the specific complexity of literature to make her point that an approach based on token frequencies can never be enough for such a complex subject matter. But she actually merges two aspects into one, which people in CLS usually treat separately and for good reasons. The first aspect is a theory of a phenomenon. The second aspect is the formal model of an indicator, which is used to test hypotheses derived from the theory and which are assumed to be directly or indirectly related to the phenomenon. The indicator will tell the researcher something about the phenomenon, if the indicator is well-chosen. But these indicators don't represent the phenomena in their full complexity. They are just one single aspect. If for example one theory assumes that authorship is solely a discursive phenomenon and another theory assumes that authorship might be even more complex involving discursive practices and similar distributions of words in texts, then a simple stylometric test can produce results which are not easily brought into agreement with the first theory. This test does not model the complexity of authorship, but it doesn't have to in order to achieve its goal.

A theory of the research process in CLS doesn't exist yet, but it will resemble in many aspects similar work done in the social sciences, where the difference between theory, hypothesis and indicator or variable, which can be measured, is widely used. In most introductions into statistics, the description of the research process is more simplified: there is a theory and a hypothesis derived from the theory. So the theory could be that a specific drug influences blood pressure and the hypotheses which will be tested look like: H0: taking the drug has no effect, H1: taking the drug reduces blood pressure. Handbooks then usually concentrate on the setup of a randomized experiment with control groups etc. They don't talk about the difficulties of measuring blood pressure reliably by using either the auscultatory or the oscillometric method. Blood pressure is a rather complex phenomenon and the oscillometric method doesn't represent all of its aspects, but uses a very specific aspect, the oscillations of a cuff pressure. Da's criticism about the use of word frequencies in CLS is basically claiming we cannot use a method because it doesn't represent the complexity of the measured phenomenon in total. As I hope the above examples show, this general claim doesn't adequately represent the complexity of research design in quantitative studies.

To take another example more related to the field, there are many differences between genres like romance and science fiction, but you can use word frequencies to distinguish between them with a very high reliability. These results are quite robust in a statistical sense. If the question, you are interested in is how the prevalence of romance and science fiction changes over time in a set of books, word frequencies may be enough to answer it. If you want to know the difference between gender representations in these genres, word frequencies may not be enough. To assume that word frequencies cannot be the basis for any kind of research question one could ask in relation to literature, because literature is too complex, is simply false. One cannot answer this question in general; you have to look at each research question and design separately: Is the chosen indicator enough to answer the research question? The task is made more difficult by our lack of knowledge in this area. Those of us in the field of CLS are also researching which kind of indicators yield robust information about which literary phenomena. Some answers will be generic, for example the use of function words for stylometric purposes, and some will be quite individually tailored to one specific question about some group of texts written at a specific place and time.

There is another variant of Da's argument also implying that literature is too complex for quantitative methods: "To look for homologies in literature, CLS must eliminate much of high-dimensional data and determine the top drivers of statistically significant variation. This always involves a significant loss of information; the question is whether that loss of information matters." (620f.) As so often in this text, the question is only rhetorical, though it actually is a real question, one which can only be answered empirically in the context of specific research questions. We already know that some questions can be answered reliably with simple representations of texts, but we don't yet know the limitations of these approaches. Da thinks she can deduce from her knowledge about the working of a procedure like principal component analysis (PCA) that it can never be used with literary texts and yield usable results. Again, she is empirically wrong, this time shown by the work of researchers like John Burrows or Hugh Craig, who used PCA rather successfully in the context of authorship attribution. It would be really hard, if not impossible, for a layman to understand from her description, how PCA works. But this is not important for her argument, because the aim is to contribute to the main theme of her text, i.e., that literature is too complex for this kind of method:

> It is one thing to statistically identify the shared drivers of a medical illness and another to say that the difference between Immanuel Kant's third critique and G. W. F. Hegel's Lectures on Aesthetics can be captured in two or three numbers derived from their overlap on two or three vocabulary lists. (621)

Da assumes again, that this application of quantitative analysis is supposed to model the whole complexity of the phenomenon. Certainly, 'the' difference cannot be captured this way, but 'a' difference can, and for some questions this will be enough.

There is a puzzling statement in the more general part of her text which shows an idiosyncratic understanding of data science: "Quantitative analysis, in the strictest sense of that concept, is usually absent in this work." (605) She doesn't explain what she thinks 'quantitative analysis, in the strictest sense of that concept' is—and it is not inference statistics, because hypothesis testing is the second item on her list of missing things in CLS—, but it is easy to show that this is not a widely shared view. Exploratory data analysis is an approach to quantitative analysis which goes back to the statistician John Tukey in the 1970s (Tukey 1977), but has gained more and more traction the more data and computing power is available. It is especially useful for detecting patterns in large collections of data. If you look at some of the recent introductions into data science like (Vanderplas 2016), typical methods of exploratory data analysis like clustering or data visualization are usually explained in depth. There is an ongoing discussion in the field about the role of exploratory data analysis in relation to confirmatory data analysis. Tukey states "Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone—as the first step." (Tukey 3) Now that large portions of a given population are in some cases available, descriptive methods have a different role to play. It isn't a resolved question what role exploratory data analysis has in the research process today, and how to make sure that the patterns, which are discovered in visualizations or by similar means, are valid and robust. Additionally the relation of traditional statistical approaches and what Breiman calls algorithmic modeling is under discussion (Breiman 2001, Underwood 2018). It is an open question how to integrate statistical methods organized around the concept of falsification with machine-learning approaches organized around the concept of optimization into one common research design framework. But this is very far from Da's claim, that 'this work' is not quantitative analysis.

All in all, Da's professed attitude that she is only helping editors in literary studies to make informed decisions is not convincing. If you look, for example, at her 'Suggested Guidelines for Reviewing CLS Manuscripts' in the online appendix of her article, she raises the bar so high for any journal editor willing to publish CLS papers by demanding a replication with the help of a programmer, a statistician and a literary historian, that no journal will have the means to go to that trouble. The guidelines of a journal like *Science* require an evaluation of the data as essential, while a software review is not required nor is replication (Science 2019). Again, the question how to evaluate data science is a real issue, and in an ideal world where research has unlimited funding specialized journals would have that combination of expertise. But as Da wields these requirements, they are just supposed to make sure that no CLS article is published in a journal of literary studies. Nan Z. Da touches some interesting questions, but because she is so bent on proving that CLS is a failed endeavor in general, she has to act as if she already has all the answers.

Selection Bias and Coda

Da calls her paper a "Computational Case against Computational Literary Studies", but if you look at her paper from a data science perspective, the research design is quite questionable, as others have also pointed out (Piper 2019). The main problem is her data. She has a sample of eight studies, which would seemingly allow her to judge an international field with many works in the last thirty to forty years. An online bibliography of the field, which Christof Schoech assembled with the help of many in the field, has more than 3,400 entries (Schoech et al. 2016). Eight replication studies is indeed a lot of work, and I am certain the field will learn something in the long run from this endeavour. But as Da has pointed out, this is of no interest to her: "It is not a method paper" (Da 2019b). She wants to show to literary studies that CLS is and will be flawed. Her statements, based on this very small sample, condemn the whole field, without any limitations, in the strongest words.

Eight data points is an extremely weak basis for that. But not only is the size of her sample a problem, the selection of the sample is too. The only thing we hear about the selection criteria and the process is that the papers were "chosen for their prominent placement, for their representativeness, and for the willingness of authors to share data and scripts or at least parts of them" (604). In statistics representativeness is approximated by a randomized selection of the cases, not by the impression of a researcher that these instances would support her hypothesis nicely.

From a European perspective the bias in Da's selection of articles in the field of Computational Literary Studies is rather obvious. It aims to talk about the field in general, but only quotes North-American scholars, especially those who have published in one journal. What about Karina van Dalen, Maciej Eder, Mike Kestemont, Jan Rybicki, Christof Schöch? —to mention just a few. All of them have published extensively in English. If you start to do research on CLS, it is difficult not to come across the journal *Digital Scholarship in the Humanities*, the flagship journal of the Association of the Digital Humanities Organisations, where a number of CLS research from all continents have been published. All of this is missing in her selection.

Additionally the bias is also distorting the image of CLS by excluding those approaches in Europe which also belong to the field but are quite different to the examples she quotes. There is for example the work of Karina van Dalen and her group in the Netherlands, who are looking into factors which make it more probably that people view something as good literature by combining methods from social science and CLS (Koolen 2018, Riddell 2018, van Cranenburg 2019). Or Evelyn Gius and Christof Meister in Germany, who are interested in robust markup schemas for narrative phenomena and how to support scholars in annotating contradictory, unclear or vague information (Gius 2017, Meister 2012). Or the work done by Maciej Eder and Jan Rybicki to evaluate and apply stylometric methods on small and large text collections (Eder 2015, Rybicki 2011) or the work done by Mike Kestemont on applying cutting edge digital methods like deep learning to languages with limited resources like Latin (Kestemont 2017) or the studies of my research group, which is adapting NLP tools to the domain of literary texts by creating manually annotated corpora of literary texts, for example for character references or speech rendering (Krug 2018). You will also find more studies about the robustness of specific quantitative methods in Europe (Eder 2015, Evert 2018, Schöch 2018). So the politically motivated bias of her selection, which has in itself a depressing effect on European readers in the current political climate, has also resulted in a seriously distorted image of the field.

The journal, which published Nan Z. Da's essay offered some of the authors, which she had criticized, the possibility for statements and she in turn answered them. Three points are of interest in her answer: 1) She emphasizes that she didn't write her text for CLS but for literary scholars and editors. In my opinion the whole structure of her paper and the selection of the examples are only understandable under this perspective. 2) She seems to be confused about what she really said in her original paper: "It seems unobjectionable that quantitative methods and nonquantitative methods might work in tandem. My paper is simply saying: that may be true in theory but it falls short in practice." (Da 2019b) As I have shown above, she does say much more. She repeatedly makes the point that quantitative methods cannot be applied to literary texts in general and not only in some specific cases. 3) She changed the outcome of her paper. In the original paper she was very clear in her statement: "the papers I study divide into no-result papers—those that haven't statistically shown us anything—and papers that do produce results but that are wrong." (605) In her later statement she modified this to: "First, there is statistically rigorous work that cannot actually answer the question it sets out to answer or *doesn't ask an interesting question at all*. Second, there is work that seems to deliver interesting results but is either nonrobust or logically confused." (Nan Z. Da: Argument, 2019, my emphasis, F.J.) With the addition of "doesn't ask an interesting question at all" she changed the first group from 'no-result papers' to 'no-result or not an interesting question'. Well, 'interesting' is a very loose and quite subjective category. This loophole is probably made necessary by the fact that in at least two cases her criticism was wrong, that is, the statistics are fine and they do show what the author of the paper intended to show. If we look at the topics of these two studies, Ted Underwood's paper on genre and Hugh Craig's book chapter on co-authorship of Shakespeare's plays, it is clear that they treat questions which have been discussed extensively in non-quantitative studies. It seems to indicate that what is not interesting to Da may be interesting to many others. Hopefully this is also true for the literary scholars and editors whom her study in resentment tried to convince of the opposite.

References

Breiman 2001: Leo Breiman: Statistical Modeling: The Two Cultures. In: Statistical Science 16,3 (2001), 199-231.

Da 2019a: Nan Z. Da: The Computational Case against Computational Literary Studies. In: Critical Inquiry 45 (Spring 2019), 601-639.

Eder 2015: Eder, Maciej Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities* 30**,**2 (2015): 167-182

Evert 2017: Stefan Evert Thomas Proisl Fotis Jannidis Isabella Reger Steffen Pielström Christof Schöch Thorsten Vitt: Understanding and explaining Delta measures for authorship attribution. In: Digital Scholarship in the Humanities 32, suppl_2 (2017), ii4-ii16.

Gius 2017: Evelyn Gius, Janina Jacke: The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis. IJHAC 11(2): 233-254 (2017).

Kestemont 2017: Mike Kestemont and Jeroen de Gussem: "Integrated Sequence Tagging for Medieval Latin Using Deep Representation Learning" In: Journal of Data Mining & Digital Humanities (2017) https://jdmdh.episciences.org/3835

Koolen 2018: Koolen, Cornelia W.: Reading beyond the Female. The relationship between perception of author gender and literary quality. Amsterdam 2018.

Krug 2019: Markus Krug, Frank Puppe, Isabella Reger, Lukas Weimer, Luisa Macharowsky, Stephan Feldhaus, Fotis Jannidis: Description of a Corpus of Character References in German Novels - DROC [Deutsches ROman Corpus]. DARIAH-DE Working Papers Nr. 27. Göttingen: DARIAH-DE, 2018. URN: urn:nbn:de:gbv:7-dariah-2018-2-9

Meister 2012: Meister, Jan Christof: Crowd sourcing "true meaning". A collaborative markup approach to textual interpretation." In: Willard McCarty, Marylin Deegan (eds.), Collaborative Research in the Digital Humanities. Festschrift for Harold Short (Ashgate Publishers) 2012, 105-122.

Piper 2019: Andrew Piper, "Do We Know What We Are Doing?" *Journal of Cultural Analytics*. April 1, 2019.

Rybicki 2011: Jan Rybicki and Maciej Eder: Deeper Delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing*, 26,3 (2011): 315-21

Science 2019: Instructions for Reviewers of Research Articles. Online: https://www.sciencemag.org/sites/default/files/RAinstr13.pdf (19.04.2019)

Schöch 2016: Christof Schöch: Stylometry Bibliography. 2016ff. https://www.zotero.org/groups/643516/stylometry_bibliography?

Schöch 2018: Christof Schöch: "Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie." In: Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven, edited by Toni Bernhart, Sandra Richter, Marcus Lepper, Marcus Willand, and Andrea Albrecht, 77-94. Berlin: de Gruyter, 2018.

Tukey 1977: John W. Tukey: Exploratory Data Analysis. Reading, Mass. 1977.

Underwood 2018: Ted Underwood: Algorithmic Modeling: Or, Modeling Data We Do Not Yet Understand. In Julia Flanders & Fotis Jannidis (eds.): *The Shape of Data in Digital Humanities: Modeling Texts and Text-based Resources*. New York: Routledge 2018, 250-263.

van Cranenberg 2019: Andreas van Cranenburgh, Karina van Dalen-Oskam,
Joris van Zundert: Vector space explorations of literary language. In: Language Resources & Evaluation (2019). https://doi.org/10.1007/s10579-018-09442-4

van Dalen-Oskam 2018: Alan Riddell & Karina van Dalen-Oskam: Readers and their roles: Evidence from readers of contemporary fiction in the Netherlands. PLoS One 13, 7 (2018)