Repeating and Repeatable: Distant Reading between Past and Future

(Christof Schöch, Budapest, Sept. 25, 2019)

1. Introduction

(A) Opening: explanation of my title

I'd like to start with a brief explanation of my title, before diving into the talk.

- In what follows, I intend to propose (and to some extend defend) a specific mode of practicing Distant Reading research, a mode that I describe as "repeating and repeatable".
- By "Distant Reading research", I mean not digital scholarly editing, not so much computationally-supported close reading, but primarily quantitative methods used for investigations of collections of literary texts. So a more descriptive, but less integrative term may be “Computational Literary Studies”.
- The principle of repeating and repeatable research, however, is much more general than that. In fact, this is an open question for me: what is specific, in this respect, to Distant Reading research?
- The mode of research I'd like to describe is (first) "repeating", in the sense that it actively seeks to link its own data, its questions or hypotheses, and its methods, to research practiced and published earlier. With the explicit aim to approximate an earlier study, but conscious also of the fact that there is no such thing as identical repetition.
- In many cases, crucially, this also means that the earlier research one is trying to repeat or “re-enact” has been practiced within the non-digital, or at least in the “non-computational”, paradigm.
- Second, this mode of research is "repeatable", in the sense that it makes all the efforts it can to provide to others the data, code, and explanatory information that make it possible for others, at a later point in time, to perform the same (or very similar) research again.
- In this way, this kind of research is located between past and future: a (never identical) re-enactment of past research, and an invitation for (never identical) further re-enactments in the future.
- This is done with the conviction, or at least in the hope, that this cycle of repetitions is not a sterile treading in the same place, but a productive, insightful upwards spiral, so to speak.

The aims of my talk

- The structure of my talk is quite simple: (SLIDE)
    - I will try to motivate and further explain my topic
    - I will propose a typology of repeating and repeatable research
    - I will discuss several case studies of this kind of "repeating and repeatable" research, mostly from the domain of Distant Reading.
    - And I will try to explain the challenges and benefits of this mode of research

(B) Motivating the Topic

- There are several factors that appear to make “repeating and repeatable research” useful today

First motivation: Fears of a disconnect between D and H

- The first motivating factor is an ongoing discourse expressing fears of a disconnect or a mismatch between the Literary Studies, on the one hand, and Digital Humanities, on the other.
- One of the main reasons for this fear is the assumption that literary texts and computational methods are simply not a suitable pairing (SLIDE): literary texts are complex, semiotic, non-deterministic, contextual artefacts and what is important about them cannot be captured by counting words.
- Inversely, this means that because DRR relies on counting surface features, it can only ask and answer a specific set of questions (like authorship attribution) but is unable to make a meaningful contribution to established, qualitative research in literary studies.
- I don’t agree with this, of course; and I see the paradigm of repeating research, in particular, with its focus on modeling and operationalization, as a useful bridge between established, qualitative research on the one hand, and computational or quantitative research, on the other.

Second motivation: reproducibility crisis

- The second motivation is the so-called "reproducibility crisis" in social psychology, medicine and the life sciences, particularly. (SLIDE)
- The term has appeared about five years ago and has served to highlight findings from a study conducted by the Center for Open Science in Charlottesville.
- In this metastudy, the authors attempted to replicate 100 papers from several key journals in psychology, all first published in 2008. They were able to replicate just 40% of them.
- There are many good reasons for this. It is really hard to avoid any and all so-called “questionable research practices” threatening reproducibility. And the studies were actually gathering relevant empirical data once more from scratch, not just checking the method. Still, a fundamental principle of the scientific method is questioned by these results.

Third motivation: Critical Inquiry paper

- The third motivating factor is the most timely and most closely connected to DRR: it is the recent paper by Nan Z Da in the influential journal *Critical Inquiry*, with the title: “The Computational Case Against Computational Literary Studies”.
- The author basically argues that computational, quantitative approaches are fundamentally unsuited for investigations into literary texts. And it argues that the selection of studies it looked at either had statistically solid results that were meaningless, or seemingly meaningful results that were not statistically sound.
- This paper is highly problematic and has been commented on and criticized extensively, for a large number of good reasons
- However, it also points to some real challenges for the field of DRR; notably, the difficulty to reproduce work in this field, starting with issues of access to data and code (e.g. da, 602).

Transition

- Now, I am not saying research as practiced today is not connected to the past and is not open to the future. Of course, we quote earlier research extensively and hope that our own research will be quoted just as extensively in the future.

- I'm also not saying it is all unsound. In fact, it is an open question whether a field like DRR can and should even be held to the kinds of standards relevant in the natural sciences. Different answers probably apply to different areas within DRR.
- But I think that repeating and repeatable research can be an important aspect to make sure our research is well-integrated into the research tradition and will be well-received in the future.

(C) A typology of repeating and repeatable research

Introduction

- To get a bit of an intuition of what I mean by "repeating" research, I'd like to use a typology to explain the various forms this kind of research can take, before delving into the case studies, for illustration and detail.

Explaining the typology

- The assumption behind this typology is that you can describe a given study, not fully, but fundamentally, by three aspects: by its research question or hypothesis, by the methodology it uses, and by the underlying data.
- Assuming, moreover, that these three aspects of a study can be either identical to the ones in the study that is being repeated, or different from them, this yields eight prototypical relations between previous research and repeating research. You see them all here, in this neat little typology. (SLIDE)
- Of course, these are simplifying assumptions. More information is necessary to provide a meaningfully complete description of a research project. But I have found this useful to clarify for myself the relationship between an earlier and a later study. (To some extent, it works into the future as well: if someone repeats the paper you published this year in five years, it'll be interesting to see which type of repetition it falls into.)
- For clarification, let's have a closer look at the left part of the typology, where the research question stays the same between the original and the later study.
- For example, a study that repeats an earlier study with precisely the same data, same method and same underlying question is, in my terminology, a replication study. In this case, you focus on checking the integrity and correctness of the method when applied to the data.
- Change the underlying data, and you are performing a "reproduction" study instead. In this case, you are checking whether, with another set of relevant data, you can show similar effects as in the original study.
- Change the method but keep the data, and you are doing a "re-analysis" of that data instead. With the research question remaining the same, you are now examining this question from a different methodological angle.
- Change both the method and the data, and you're doing "follow-up" research on the same research question. This is what we do most of the time, in some sense; myself included. I happen to believe this is the least useful type of research.

Limitations of the typology

- The typology seems useful to me, but I am not 100% happy with it, of course: I find it troubling, for example, that the typology assumes a binary relation of identity or difference. We know (and will see in the case studies), that there is in fact a rather fluid relation, for example, between data that is identical, almost the same, somewhat similar, rather different, etc. from

the original data. And the same is true for the methods used, particularly when you repeat research across the analog-digital divide.

2. Case studies

(A) Opening

How does it work?

- So, how does this "repeating and repeatable" research work in practice?
- I will take a bit of time to describe a few examples, before trying to derive some general conclusions from them

(A) Mendenhall repeated by Rockwell and Sinclair

- My first encounter with the idea of repeating research (as an explicit research principle) goes back to the year 2015, when Geoffrey Rockwell gave a talk titled “Replication as a way of knowing” in Würzburg.
- He presented work he had done together with Stéfan Sinclair on re-enacting a classic, very early, stylometric study by Thomas C. Mendenhall. (SLIDE)

Background

- In 1887, Thomas C. Mendenhall published an article in *Science* titled: “The Characteristic Curves of Composition”
- His fundamental idea was that you could identify the author of a text by the distribution of word lengths in their texts
- For example, this is the word length distribution plot that Mendenhall obtained for Oliver Twist (SLIDE).

The repeating study

- Stéfan Sinclair und Geoffrey Rockwell had the idea to re-implement this paper
- Their example is Mendenhall's word length distribution plot for Charles Dickens' Oliver Twist
- When Sinclair and Rockwell repeated Mendenhall’s experiment, they got pretty much the same result. (SLIDE)
- Then, they also did this for the entire novel and got a very similar distribution. (SLIDE)

Comments

- Many of the features of repeating and repeatable research are already present in this seemingly simple study.
- The starting point is a more or less famous example of early quantitative research.
- We find a close alignment in terms of the data: the digital edition Rockwell and Sinclair used (from Project Gutenberg) is probably even based on a print edition not too far away in time from the one Mendenhall used!)
- We also see a display of the advantages of the digital in terms of scale: once the code works with 1000 words, it is trivial to expand the analysis to the entire novel.
- We see efforts to approximate the earlier results, but also an attention for the slight and inevitable deviations; note that there seem to be a few words with 13 letters that Rockwell and Sinclair found but Mendenhall did not; a question of the underlying text? or of tokenization?

- The authors use this research strategy not so much as a way of checking on Mendenhall, but to better understand his approach and, in the process, also think about their current, computational methodology
- Finally, the authors understood the link between repeating and repeatable research. Therefore, they make their own research processes repeatable. In this case, their strategy is to provide a jupyter notebook that lets you easily run the code and see what is going on; almost too easily! (SLIDE)
- What Rockwell and Sinclair don't do is actually try to perform authorship attribution with this kind of data. This, in fact, turned out not to work very well. The problem is that all of these plots are actually quite similar and that the variation you do find does not correlate well with authorship.

(B) My first and simplest own attempt at this: Richaudeau on Simenon repeated by myself

Introduction

- Let's move on to another case study, this time a repeating study I did myself.
- In 1982, Francois Richaudeau published a paper on Belgian writer Georges Simenon. His focus was on sentence length distributions. (SLIDE)
- The main results of Richaudeau were the following: in Simenon's work, the average sentence length is generally quite low (about 12-13 words); and it is different for crime fiction ("Maigret novels") and psychological fiction ("romans durs"), but only slightly so. The autobiographical texts have much longer sentences. (SLIDE)

Replication

- This paper lends itself easily to being repeated, simply because it is already a proto-digital, quantitative study entirely focused on one simple surface feature.
- It also helps that the availability of Georges Simenon's work in digital form is very good, so that I was able to establish Richaudeau's somewhat idiosyncratic selection very closely.
- What my results with this corpus showed was that, yes, the resulting average sentence lengths for different parts of Simenon's work could be confirmed, generally, by the repeating study. I was also able to show that a comparison of sentence lengths for the Maigret novels versus the "romans durs" did not show a statistically significant difference. (SLIDE)
- However, Richaudeau only had 25 novels in his set, probably because his rather elaborate method of sampling sentences and establishing their length was extremely time-consuming. The data becomes much more convincing when you add more data points. (SLIDE)
- However, the study also showed that Richaudeau had been missing suitable comparison texts. His previous study on sentence lengths had been on Proust, and of course Simenon's sentences appeared short after that.
- With a large comparison corpus of novels contemporary to Simenon, it became clear that he did certainly not use unusually short sentences, but he was right there in the middle of everyone else. (SLIDE; Simenon in green)
- So sentence length alone was not what could explain Simenon's popularity.

There are several lessons here

- First: You never have the same texts, so that a pure replication (see: typology) of an earlier, non-digital study is almost impossible. However, deliberately introducing more data, like adding

500 novels for comparison, does appear to be useful even for a better understanding of the original study.

- Second: it is tricky to use exactly the same definition of word and sentence, when a paper does not make their definitions explicit. This is really a lesson for us when writing papers: make these details as explicit as possible, in prose in a methodological supplement and by publishing your code. (Make sure your prose fits your code, too! ;-)
- Third, digital methods scale admirably. however, it is easy to dismiss earlier studies for their small datasets; we have it so good now! (And providing useful datasets is so important. Withouth, we can do nothing.)
- Finally: In this particular study, making the repeating study also repeatable is quite a challenge, not least for reasons of copyright: While it is possible to publish derived data, such as sentence length statistics, I cannot publish the full novels for someone else to do the same kind of test. (Could I publish a random sample of full sentences in random order? In Trier, we're currently preparing work providing answers to such questions, together with colleagues specializing in digital copyright law.)

(C) My most serious attempt at this: repeating Spitzer on Racine

Introduction

- We now turn to another case study, this time of a stylistic analysis published by Leo Spitzer in 1928, on French classical author Racine, best known for his tragedies.

Spitzer's original study

- This study is different from Mendenhall and Richaudeau in many ways, but most importantly by the fact that it is not a quantitative study, but really a very careful, qualitative stylistic analysis and interpretation.
- Spitzer's fundamental hypothesis is that Racine is an example of what he calls the "Klassische Dämpfung". The French translation calls it "l'effet de sourdine".
- What Spitzer sees at work in Racine's style is an effort to "dampen" any strong expression of emotions, or at least some alternation between dampened and undampened passages.
- This dampening actually becomes the defining stylistic trait for Spitzer, and he traces it in close to 50 different stylistic devices across Racine's work. Spitzer describes and illustrates each of them in the paper. It's a long paper!!
- To discuss just one example: the "where that effaces boundaries". (INSERT FRENCH CITATION!)

Repetition of this study

- To repeat this study, I obtained the complete works of Racine (from theatre-classique.fr, bien sûr!).
- Then, I tried to implement a search pattern, using WordNet annotations, word lists and regular expressions, to closely model each of the 50 stylistic patterns described by Spitzer. For this, I used the corpus analysis tool TXM.
- I was successful only in 30 out of 50 patterns; the remaining patterns were simply too strongly dependent on semantics and context to be modeled in this way.
- To check the quality of the modeling, I performed a simple test for precision and recall: All resulting passages should match the pattern described by Spitzer, and all examples mentioned by Spitzer should be included in the resulting passages.
- Some results from doing this for the "boundary-effacing where" are shown here. (SLIDE)

- Doing this was highly instructive to understand how Spitzer proceeds; what is necessary to model his patterns formally; how the apparent precision of his descriptions is misleading; how context-dependent the formal patterns are that he describes; and how limited the technical means are that we currently have.
- For example, Spitzer defines the pattern involving “où” as “abstract nouns followed by où” but in fact primarily means words of emotions; an important shift when you try to model the phenomenon.

Going beyond Spitzer

- One more thing: Spitzer leaves the question of how Racine relates to other, contemporary authors of tragedies, relatively open. For him, Racine is “a unique star” and would be diminished by comparison.
- But the title “Die klassische Dämpfung bei Racine” implies that there is such a things as a more generally classical “effet de sourdine” that can not only be observed in Racine, albeit in a privileged manner.
- In any case: testing the prevalence of the instances of the 30 stylistic patterns against a comparison corpus of contemporary tragedies shows that only 3 out of the 30 patterns are in fact significantly over-represented in Racine as opposed to the contemporaries. (SLIDE) That’s very few! And two are significantly under-represented!

What are the lessons from this study?

- First of all: Qualitative, non-digital studies, even so clearly stylistic, formal and detailed ones as Spitzer’s work, are really not meant to be repeated. It is tough!
- Also, the relation between the original study and the repeating study is much less binary as my typology supposes. The data and method I used, are approximations of the ones Spitzer used, but cannot usefully be described as either “same” or “different”. In fact, the longer I worked with the Spitzer article, the more my research aims deviated from the original.
- In working hard to approximate an earlier study, and failing to do so, you really learn a lot about the specificities of both the original study, and your own study, in terms of data and methods.
- In this study, the challenge of how to make sure my own work is easily repeatable by others, in the future, is very clear. Not for legal reasons, this time, but for rather practical reasons: how can I document my queries inside TXM, and the resulting list of relevant passages, in a standardized, accessible, and formalized way? (As far as I know, this is an unsolved issue in this particular context).

3. Conclusions

- To conclude: this is all well and good, and I can assure you that doing this kind of work is really exciting and instructive.
- But there are challenges as well as benefits to it.

(A) Challenges

General challenge: Potential (or apparent) conflict with disciplinary values

- It seems clear that my focus on "repeating and repeatable" research may be perceived to be at odds with the ways in which we usually ascribe value to research.
- In order to get funding, to be accepted at a conference, or to be published in a journal, research usually needs to be original, innovative, ground-breaking, going beyond the state-of-the-art; "excellent", in a word.
- The kind of research I advocate for, on the other hand, is fundamentally concerned with repeating research that has been done decades or at least years ago by others already; and that aspires to be repeated by others in the future. Can such research be said to be excellent in this sense, or to foster excellent research?
- Of course it can; in fact, we need reproducibility simply as quality assurance; and we need repeating earlier research. I think that guaranteeing the continuity, over time, of the disciplinary context of our work is a worthwhile endeavor.

Challenges specific to work in Digital Literary Studies

- One challenge that is specific to work in Digital Literary Studies has to do with the data: In contrast to the sciences, we cannot simply repeat an empirical study with a different set of subjects or by taking measures again in identical circumstances.
- This is simply due to a lack of available data: often, studies have already used all the data that can be obtained and that is relevant for the question. There aren't any other Racine plays than the ones Spitzer analyzed, for example. Take another author, and everything will be different; of course!
- Another challenge is the following: It seems clear to me now that pure replication is the least interesting case. Only once you depart from identical question, data and code or method, by introducing slight deviations from the original study, does real insight occur. The epistemological value lies not in replication, but in the subtle departure from the original study design. But then, how do you find the right balance between faithfulness to the original and productive departure from it?
- In my experience, this may be a false problem: it is difficult enough to approximate an earlier study, so deviation will occur even if you try to avoid it; at least in the case of repeating non-digital studies, where the data needs to be captured and to code-based implementation written from scratch. It is more an issue with recent, fully-digital studies.
- Also, and rather in the perspective of repeatable studies: if this kind of departure from the original is the most productive way of re-using earlier data and code, how can we best enable this? For example, making choices in the code very explicit and coding the alternatives as well, even if we ourselves believe our ultimate choices were the right ones.

Benefits of doing this kind of research

- Despite these challenges, there are plenty of clear benefits to this kind of research.
- First of all, you understand previous research much better when you try to re-implement it. You recognize its strengths, and you see its limitations.
- Second, you also understand the strengths and limitations of your own approach much better! We can see that Digital Literary Studies is in fact very good at counting words or letters, and use this at a scale inaccessible to non-digital work.

- But we can also be challenged to push the limits of what is possible in terms of modeling of literary phenomena by research that does this using human understanding and intuition. There is still lot of work to do to improve our tools.
- By the way: I also advocate for using this kind of approach in teaching. By re-implementing prior research with students, they learn a lot more than by just reading the paper.

Final Words: An Agenda for repeating and repeatable DRR

- So, my final words:
- I do think that we should aim to make our research reproducible in the way research in the sciences aims to be reproducible, which means publishing data and code alongside any paper. DRR works empirically, and we should accept to be held by the standards of empirical research.
- However, this is only a positive side-effect of the "repeating and repeatable" mode of research in DRR that I've tried to illustrate and defend.
- What really counts is two things:
    - Practicing "repeating and repeatable" research connects your research strongly and explicitly with both the past and the future of your field, ensuring diachronic disciplinary coherence.
    - And "repeating and repeatable research" permits deep methodological insights, both into earlier work and into our own work.
    - In a nutshell, "repeating and repeatable research" is a useful learning experience. Thank you.