TECHNISCHE UNIVERSITÄT DARMSTADT
**DIGITAL HUMANITIES COOPERATION**
UNIVERSITÄT KONSTANZ

# Thomas Weitin

## **Thinking slowly.**
## Reading Literature in the Aftermath of Big Data

LitLab

Pamphlet #1
**March 2015**

Ed. Thomas Weitin

FORSCHUNGSSTELLE
DIGITAL HUMANITIES CENTER
UNIVERSITÄT KONSTANZ

Thomas Weitin

# Thinking slowly

## Reading Literature in the Aftermath of Big Data

## Abstract

Literatur schult mit dem ‚langsamen Denken' eine Kompetenz, die es im Zeitalter digitaler und statistischer Verfahren zur Textanalyse unbedingt zu bewahren gilt. Gleichzeitig bietet der operationalisierungsorientierte Ansatz, den quantitative Verfahren verlangen, die Gelegenheit, methodische Selbstverständlichkeiten wie den verifikationsorientierten Denkstil der Hermeneutik zu hinterfragen. Ohne diese Kritik lassen sich Text- und Datenanalyse nicht erfolgreich kombinieren.

Literature is a training in slow thinking which should be cherished in the Digital Age. At the same time the operationalization of textual analysis is not only a necessity of quantitative approaches but also a chance to rethink certain habits of hermeneutics such as the longing for verification. Without that textual and data analysis can't be combined successfully.

Bibliographical information of the German National Library:
The German National Library lists this publication in the German national bibliography. It is registered in the data base of journals (ZDB) and in the international ISSN portal. Bibliographical details can be found online at http://dnb.d-nb.de.



ISSN: 2628-4537

Ed. by Thomas Weitin

*Thomas Weitin*

# Thinking slowly

## Reading Literature in the Aftermath of Big Data

We literary scholars subsist on reading! We read and write about it and we instruct others who read and then, in turn, go on to teach others to read. We are the pros in this field. We read, understand, and analyze literature in a methodical way that no layperson can readily imitate. We can even rather swiftly interpret difficult texts and produce a well-founded version of that reading that stands the test of time.

What does this competency mean in light of that overpowering reader who always already reads along while we are still searching for words? The search-engine company from Mountain View, California that has grown up under the sign of the California bear has for some time now come to resemble the fencing bear from Kleist's *On the Marionette Theater* who not only parries all the thrusts of his human opponent but does so as if he could "read [..his] soul,"[1] always knowing in advance which direction the human being will turn. Has the Kleistian fantasy of the super-reader become reality? Common etymological roots are discernable [in the German, at least] for a "reading" that is hermeneutic and the "culling," in the sense of "reading out," that constitutes the gathering and searching that we associate with the algorithmic activities of digitally recorded hardware. But how do they relate epistemically to one another?

In what follows, I would like to do no more than consolidate this question. In order to likewise account for overstrained anxieties and unrealistic expectations, I take a look at a series of distinctions that are frequently made to stir up such anxieties and expectations. There is, for one, the distinction fast/slow: Which procedures are fast, and which are slow? Who operates superficially, and who goes into depth? And in what ways do we attain new knowledge with and about texts? How does hermeneutic *close reading* relate to *distant reading*? For our colleague from Stanford, Franco Moretti, this battle cry includes the culling of large amounts of texts according to their formal features and with the aid of statistical procedures.[2] As a provocative refinement of different methodological options, however, this distinction has meanwhile done its job. As an alternative, I will explain why the integrative concept of *scalable reading* is

[1] Heinrich von Kleist: *Sämtliche Werke und Briefe*. Vol. 2, München: Hanser 1965, p. 345.
[2] Cf. Franco Moretti: *Distant Reading*, London, New York: Verso, p. 47 ff.

more suitable for projects that seek to combine the qualitative methods of hermeneutics and quantitative methods of statistics.[3] First, I would like to begin with the question: What does big data mean for philology? Do we want it, do we need it, or does aesthetic knowledge reside in the idea that "the small is beautiful?"

## 1. Recognizing Text

Actual big data-analyses do not yet exist in the humanities. Even the Californian super-reader bear from Google Books has only read, as of April 2013, around a good 30 million of the estimated 130 million books that humanity has created up to now.[4] Since then, of course, additional books have been scanned worldwide. And the tool that Google offers us as a reading aid, the *n gram viewer*, is eagerly used by colleagues who comply with the request on the start page to "run your own experiment." Thus did the historian of Eastern Europe, Alexander Etkind, from King's College in London, present a study on "Mourning and melancholia in putin's Russia," in which, on the basis of the frequent naming of current Russian politicians along with the name of Stalin, the equally high hit rate of central dates of Soviet history (1917, 1937, 1941, 1945), as well as the overrepresentation of historical keywords compared with current topics, he claimed to identify a widespread melancholy in a Russian society that basically lives in the past.[5] Students of mine tried to repeat the experiment with the years and with the Russian Google Books corpus used by Etkind and were initially very surprised when the graphs showed a different course as the charts published by Etkind.[6] Later, Myriam Traub (Amsterdam) indicated that besides the current corpus there is an older one from 2009—indeed, this was the cause for the different outcome. Having become skeptical, they tested just for fun the conceptual

[3]Cf. Martin Mueller: „Morgenstern's Spectacles or the Importance of Not-Reading." In: *Scalable Reading* (blog), 21.01.2013, https://scalablereading.northwestern.edu/2013/01/21/morgensterns-spectacles-or- the-importance-of-not-reading/.

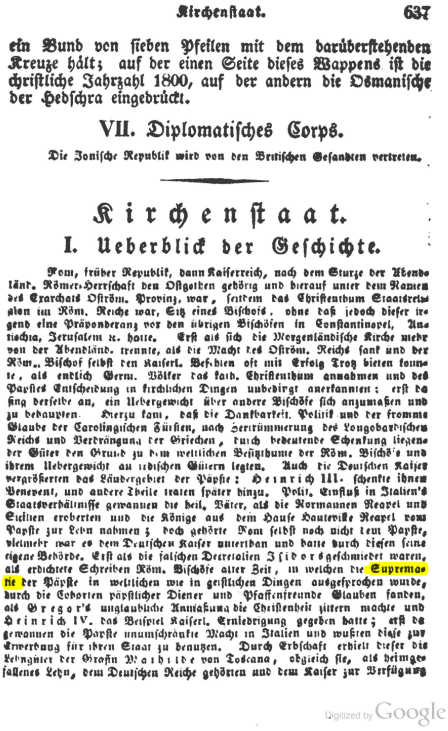
[4]Cf. Christof Schöch: "Big? Smart? Clean? Messy? Data in the Humanities." In: *Journal of Digital Humanities* 2.3 (2013), p. 8. The estimations are from Google itself from the year 2010. Schöch cites the figure according to Wikipedia, and while more current figures are not quoted today (March 2015), there is a note that scanning at American libraries has slowed. The original goal was to have scanned all the books of the world by 2020.

[5]Cf. Alexander Etkind: "Mourning and melancholia in putin's russia. An essay in Mnemonics." In: Ellen Rutten, Julie Fedor, Vera Zvereva (eds.): *Memory, Conflict and New Media. Web Wars in Post-Socialist States*, London: Routledge, 2013, p. 32-47.

[6]Katharina Herget conducted this experiment in my seminar on "Hermeneutics – Statistics – Cognition" in Summer Semester 2014 at the University of Konstanz.

career of "superman" and arrived once again at astonishing results. Alongside the predictably steep ascent of the superhero since the nineteen-forties, a slight but obvious curve already starts to arch between 1840 and 1860. An archetype?, we wondered in the seminar. Nietzsche had hardly been born and thus was ruled out of the circle of usual suspects. Ultimately, we discovered a row of words in the hit list [corpus language was set to German] that Google falsely detected as "superman", among others the *Genealogical-Historical-Statistical Almanac* for the year 1840, which in a chapter on the history of the Vatican refers to the "supremacy [German: **Suprematie**] of the Pope."


Figure 3: Screenshot of the *n gram viewers*, keyword: ‚superman'


Kirchenstaat. 637

ein Bund von ſieben Pfeilen mit dem darüberſtehenden Kreuze hält; auf der einen Seite dieſes Wappens iſt die chriſtliche Jahrzahl 1800, auf der andern die Osmaniſche der Hedſchra eingedrückt.

VII. Diplomatiſches Corps.

Die Joniſche Republik wird von den Britiſchen Geſandten vertreten.

Kirchenſtaat.

I. Ueberblick der Geſchichte.

Rom, früher Republik, dann Kaiſerreich, nach dem Sturze der Abendländ. Römer-Herrſchaft den Oſtgothen gehörig und hierauf unter dem Namen des Exarchats Oſtröm. Provinz, war, ſeitdem das Chriſtenthum Staatsreligion im Röm. Reiche war, Sitz eines Biſchofs, ohne daß jedoch dieſer irgend eine Präponderanz vor den übrigen Biſchöfen in Conſtantinopel, Antiochia, Jeruſalem ꝛc. hatte. Erſt als ſich die Morgenländiſche Kirche mehr von der Abendländ. trennte, als die Macht des Oſtröm. Reichs ſank und der Röm. Biſchof ſelbſt den Kaiſern. Beſtrien oft mit Erfolg Trotz bieten konnte, als endlich Germ. Völker das kath. Chriſtenthum annahmen und des Papſtes Entſcheidung in kirchlichen Dingen unbedingt anerkannten; erſt da fing derſelbe an, ein Uebergewicht über andere Biſchöfe ſich anzumaßen und zu behaupten. Dazu kam, daß die Dankbarkeit, Politik und der fromme Glaube der Carolingiſchen Fürſten, nach Zerrümmerung des Longobardiſchen Reichs und Verdrängung der Griechen, durch bedeutende Schenkung liegender Güter den Grund zu dem weltlichen Beſitzthume der Röm. Biſchöfe und ihrem Uebergewicht an irdiſchen Gütern legten. Auch die Deutſchen Kaiſer vergrößerten das Ländergebiet der Päpſte: Heinrich III. ſchenkte ihnen Benevent, und andere Theile traten ſpäter hinzu. Polit. Einfluß in Italien's Staatsverhältniſſe gewannen die heil. Väter, als die Normannen Neapel und Sicilien eroberten und die Könige aus dem Hauſe Hauteville Neapel vom Papſte zur Lehn nahmen; doch gehörte Rom ſelbſt noch nicht dem Papſte, vielmehr war es dem Deutſchen Kaiſer unterthan und hatte durch dieſen ſeine eigene Behörde. Erſt als die falſchen Decretalen Iſidorsgeſchmiedet waren, als erdichtete Schreiben Röm. Biſchöfe alter Zeit, in welchen die Suprematie der Päpſte in weltlichen wie in geiſtlichen Dingen ausgeſprochen wurde, durch die Cohorten päpſtlicher Diener und Pfaffenfreunde Glauben fanden, als Gregor's unglaubliche Anmaßung die Chriſtenheit zittern machte und Heinrich IV. das Beiſpiel Kaiſerl. Erniedrigung gegeben hatte; erſt da gewannen die Päpſte unumſchränkte Macht in Italien und wußten dieſe zur Erweiterung für ihren Staat zu benutzen. Durch Erbſchaft erhielt dieſer die Herrſchaft der Großn Mathilde von Toscana, obgleich ſie, als heimgefallenes Lehn, dem Deutſchen Reiche gehörten und dem Kaiſer zur Verfügung



Figure 4: Page where "supremacy" [German: "Suprematie"] was falsely recognized as "superman"

That through a technical oversight we read with Google of heroes where there are none is certainly not the problem. But it is problematic in light of the growing popularity of the *n gram viewers* in the humanities that I, as one who would hold the belief that I am conducting my own experiment, have no control over the corpus that I am investigating. It so happens that results which I produce can no longer be found again. And even when I can successfully move from the frequency diagram to every single book in the hit list, the corpus in its entirety remains a black box for me. That may not be unusual for an analysis of big data, but it does raise

the fundamental question of the heuristic status of readily producible *n gram views*. There is a danger that these data images will be misunderstood as evidence for presumptuous theses, whereas the argumentation in fact uses them to rationalize what they have "shown" but not proven. When for instance Etkind demonstrates the preoccupation of post-Soviet Russian society with the tragic events of its history through the precipitous development of the graph for "1941" (in contrast to the positive date of "1945") and connects the abrupt end of its rise with the beginning of the reign of Putin, he produces this type of questionable evidence. Proof has not yet been furnished by the observed simultaneity of childbirth and the arrival of storks. Alternative explanations are still possible.

## 2. Smart Data

Everything negative that can be said about Google in this regard turns out to be entirely different in the work with other popular tools that have been developed for quantitative text analysis in different university settings worldwide. Tools such as "AntConc" from the Waseda University in Tokyo, "Voyant" from the Canadian universities of McGill and Alberta, or "CATMA" from the University of Hamburg, are not only freely available but also transparent, in part even when it comes to the construction of the program.[7] Here I can determine for myself the size of my corpus, I can investigate individual texts and undertake corpus analyses. Full corpus-control is guaranteed. Regardless of how large of a corpus I select, whether for instance I import a particular novella, all novellas by a certain author of an epoch, or those of an entire century, I do not thereby produce big data. In place of the distinction between big or small data, another distinction becomes relevant for my procedure, namely the one between structured and unstructured data. "Plain text" that I import into a tool counts as a prime example of unstructured data. The tools of analysis can calculate word frequency from it, or the word distribution in texts as well as in corpora, they can create concordances via the Keyword-in-Context function. This can be useful for orientation in the text or corpus and it can deliver evidence for or grounds to reject a hypothesis. However, the tools are most effective for work

[7] The tools can be accessed at the following URLs: "AntConc" (http://www.laurenceanthony.net/software/antconc/), "Voyant" (http://voyant-tools.org), "CATMA" (http://www.catma.de). In the case of "CATMA," one can directly see how the program is built: http://www.catma.de/technology.

with structured data, that is, with text which is annotated or furnished with markups. Linguistic features can in part already be marked automatically and made legible, though work with literary texts in particular will always require an enormous "manual" effort to produce *smart data* out of unstructured text data for an analysis. Whoever undertakes that must have strong nerves, because no one can guarantee that the entire effort will pay off and that in the end the annotated corpus will in fact help to support an original hypothesis. Smart data compensates for this, however, by the fact that the structures that I investigate are explicit and clear. This is not the case with big data, which presents a significant problem. Christof Schöch has referred to the ways in which big data represents something fundamentally different for the humanities than it does for the natural sciences or the economy.[8] For example, the question central for economic analysis concerning the speed and density of data streams has not played a role up to now in inquiries in the humanities. With a view to future inquiries and our methodological self-understanding, however, the characteristics of big data compared to those of smart data are significant. Smart data still requires proper manual labor and is thus not the stuff about which computer scientists dream. I can more quickly start to calculate with big data, but because large quantities of data in the form of text are much more unstructured, I am entirely reliant upon statistical methods and – this is crucial – their visualizations, such as the results of cluster or principle component analyses (PCA). "Big data requires visualization to even start understanding its possible structures."[9] This can become a problem when humanists untrained in statistics draw premature conclusions based on the often aesthetic images of statistical analyses. In the worst case, they sit, as with Google, in front of a black box without even letting it worry them a bit.

## 3. Methodological Clichés

Against this adumbrated backdrop, it is time for a brief critique of the methodological clichés with which both opponents and proponents of the quantitative analysis of literature try to outdo one another. The common prejudice that quantitative analyses "simply [let] algorithms run over

[8]Cf. Christof Schöch [see note 4]: "Big? Smart? Clean? Messy?", p. 7.
[9]Ibid., p.10.

texts," disregards the intellectual effort of the "operationalization"[10] of research questions and deprives itself of the productive effect that operationalization-oriented thought can also have for conventional readings that are thereby beholden to the precision and traceability of every step. We can assume that the cause for the aversion has much to do with a style of thought that is widely habituated in literary studies and that cannot get accustomed to the fact that the present methodological innovation in the field cannot be followed by reading yet another book in order to enrich one's reading with another new theoretical concept. Quantitative analyses demand a high expenditure of time until one is able to even use the corresponding tools and environments of analysis. What is easy to use – see the *n gram viewer* – might not yield reliable results.

A tool like "AntConc" offers a good compromise. It is rather easily accessible and conveys fundamental quantitative parameters that can be scaled elegantly between a single text, groups of texts, and larger corpora. Entirely different dimensions are introduced, however, by the collaborative analysis environment "RStudio" that is constantly updated with new *packages* and can do everything that the tools can but also offers possibilities for "classical" statistical visualizations (PCA, cluster), for machine-learning-based text analysis as well as for network analyses and topic modeling. Here I not only have full corpus control, depending on the input, but as an author of R-scripts can also work directly on the operationalization of my research question, for which of course I must first learn the programming language. In addition, for texts from the nineteenth century in particular, there is the extremely laborious text recognition using an OCR software whose results (above all with Fraktur, or Gothic script) have to be corrected by hand. It is possible that a normalization of the orthography might be required.

In light of such an investment of time for merely the preparation of quantitative analyses (whose operationalization, implementation and assessment is moreover accompanied by considerations of literary history as well as by hermeneutic and above all structural observations of the text), the second common objection sets its sights on the uncertain prospects and the often straightforward yield of the analyses. Do they not in the end merely confirm with much technical effort that which we already know? That adjectives for emotionality accumulate con-

[10] Cf. Franco Moretti: "Operationalizing: or, the function of measurement in modern literary theory." *Stanford Literary Lab. Pamphlet* 6 (2013): 1-13.

spicuously in the German-language Age of Sentiment or that Hemingway wrote with a more journalistic style than Joyce, that in the genre of comedy the topic of "pleasure" is strongly represented – does that have to be quantitatively confirmed? The knee-jerk answer of many literary scholars is: of course not, what is the point? And this reaction comes so self-evidently, because the original new thesis or observation traditionally counts in the discipline for more than the consolidation of an already known result through an alternative method.[11] But because the methodological upheaval in the aftermath of digital techniques of analysis proves to be time-consuming and tedious, it is worth changing the traditional style of thought to the extent that we make ourselves familiar with these techniques in a manner that is both consolidating *and* expectant of new insights.

There are encouraging indications that this path will promise success. Through the assessment of metadata, Mathew Jockers was able to show that the literary historiography of *Irish-American fiction* is based on a small canon out of which narratives of the boom and crisis of this literature, as well as of their themes and settings, have been constructed.[12] If the latter are compared, as Jockers did, with the statistics of the metadata, if in other words the absolute publication numbers, the information about the place and year of publication as well as the name and gender of the author is assessed within the framework of the entire corpus, it then becomes clear that the narratives of literary historiography use a very particular cross-section in order to tell the history of the putative whole that they in truth do not even have on their radar. In this case, the canon consists of male *Irish-American* authors from the American East Coast whose works are set in rural areas. Not accounted for, but counted by Jockers are the majority of female authors who went to the West Coast and wrote primarily about urban topics. If one factors them into the equation, the extant narratives of the rise and fall of *Irish-American fiction* and their identity-constituting themes appear as inadmissible generalizations.

Studies like this one show that statistical data analysis can indeed correct the findings

[11] In her critique of "distant reading" in the *New York Times* from 26 June 2011, Kathryn Schulz criticized Moretti's approach for merely confirming existing knowledge and hence being banal. Jockers identifies in this critique a disdain for the literary object: "Why should further confirmation of a point of speculation engender a negative response? If the matter at hand were not literary, if it were global warming, for example, and new evidence confirmed a particular 'interpretation' or thesis, surely this would not cause a thousand scientists to collectively sigh and say, 'Duh.' " (Matthew Jockers: *Macroanalysis: Digital Methods & Literary History*, Urbana, Chicago and Springfield: University of Illinois Press, 2013, p. 31)

[12] Cf. Ibid., p. 35-62.

of a literary history produced by individual readings. The question is what methodological conclusions do we want to draw from this? A certain tendency found in Moretti and Jockers seems questionable in this respect, because they resort to methodological clichés. Most likely due to the fact that quantitative procedures can still count on the spontaneous aversion of most literary scholars, the popular champions of digital and statistical methods cultivate for their part a disparaging prejudice against what they refer to in the American tradition as *close reading*. The corroboration of hypotheses through individual readings that work closely with the text is regarded by them as a merely "anecdotal" procedure that arises from a "disciplinary habit of thinking small."[13] Literary history, they claim, can only be written according to the law of the large number, that is, using statistical procedures based on big data: "Cherry-picking of evidence in support of a broad hypothesis seems inevitable in the close-reading scholarly tradition."[14] "Close reading is not only impractical as a means of evidence gathering in the digital library, but big data render it totally inappropriate as a method of studying literary history."[15] No doubt to provoke the habits of innovation of his colleagues in literary studies, Moretti exaggerates the critique of method with the lapidary sentence: "[N]o one has ever found a method by just reading more texts."[16]

The corpora of literary history, so is the claim, are simply too gigantic to be read through individually. This argument is repeated often. It emerges in tandem with the methodological cliché of the *close reader* who picks out of a text that which tastes best to his or her respective theoretical predilection. The battlefront drawn here between small versus big data, *close reading* versus *distant reading* supplies a very imprecise atlas of the contemporary research landscape. Procedures taking place on the middle level, mid-size copora processed as *smart data*, are hardly appreciated. And if anything, big data is entirely overestimated. Above all, this battlefront is unproductive for the development of future projects, because the step from consolidation to innovation can only succeed if it synergistically combines our hermeneutical reading with the digital analyses of the computer in concrete analyses. Jockers seems to be on the right track when in spite of the rhetoric of big data he writes: "The two *scales* of analysis, therefore, should

[13]Ibid., p. 46, 16.
[14]Ibid., p. 47.
[15]Ibid., p. 7.
[16]Franco Moretti [see note 2]: *Distant Reading*, p. 46.

and need to coexist."[17] Yet he does not pursue the keyword "scale," which he already has on his screen, and positions himself as a purely *distant reader*, whereas Moretti once again has started arguing strongly from a standpoint oriented toward individual texts in his more recent conceptual works.

Martin Mueller proposes "scalable reading"[18] in the sense of a guiding concept, although it is one guided by a theoretical insight that has been known for a while but has once again gained importance in the current situation. Just about all texts that serve as a material basis for hermeneutic reading are available to us in a form that is different than their original form of production, without us being always consciously aware of this while reading. Whoever reads the Oxford-Classics edition of the *Odyssey* (according to Mueller's example), is in fact already a *distant reader* of the songs ascribed to Homer. And precisely the *distant reader* in the conventional sense, who generates data and visualizations through digital analyses, must understand and interpret them.[19] *Scalable reading* does not (only) mean however that *close* and *distant reading* mutually inform one another methodologically (this analogy might entice us to underestimate the differences between the interpretation of text and data); it stands for an integrated understanding of all acts of reading and analysis. As a rule, and this is Mueller's old observation tethered with new relevance, we read and analyze texts in the form of "surrogates"[20] the orally transmitted epic fixed in written form, drama in a critical-historical edition with commentary, the novel edited with references to all variations, as well as the novella in txt-format as a concordance plot of certain keywords whose "keyness" I calculate (with "AntConc") in relation to a reference corpus, or the eighty-six novellas from Paul Heyse's *Treasure Trove of German Novellas* rendered on the basis of xml-files in a word-frequency-profile-based dendrogram or in the Topic Modeling of their semantic probabilities (both within "RStudio"). The concept of *scalable reading* considers all of these forms of reading and analysis as based on

[17] Matthew Jockers [see note 11], *Macroanalysis*, p. 9; emphases TW.

[18] Martin Mueller [see note 3].

[19] "Whether derived by machine or through hours in the archive, the data through which our literary arguments are built will always require the careful and imaginative scrutiny of the scholar. There will always be a movement from facts to interpretation of facts. The computer is a tool that assists in the identification and compilation of evidence. We must, in turn, interpret and explain that derivative data." (Matthew Jockers [see note 11], *Macroanalysis*, p. 30)

[20] "Our typical encounter with a text is through surrogate — setting aside whether there is an original in the first place." (Martin Mueller [see note 3])

a respective surrogate whose characteristic mediality offers specific possibilities of knowledge.[21] This ranges from texts written on paper to digital surrogates that, using word-frequency lists as their basis, convey *z-score*-matrices and *delta scores* or transform the continuous text into a *bag of words* in order to be able to model it. We are thus dealing with a broad "scale" of surrogates that can be put into operation as a continuum, as for instance when with collections such as Heyse's *Treasure Trove of Novellas* I juxtapose the results of individual readings with those of digital corpus analysis.

Like any other concept, *scalable reading* can only stand the test of time in concrete research projects, yet a more fundamental systematic advantage is that it makes a continuum of qualitative and quantitative methods conceivable[22] where previously the understanding of "traditional hermeneutics" on the one side and "new techniques of analysis" on the other, despite all protestations to the contrary, has not really been interested in productive connections.

## 4. *Scalable Reading*

Although theoretical concepts should prove themselves primarily in concrete research projects, they also take on a general exploratory function during the phase in which the humanities acquaint themselves in grand style for the first time with the procedures of *digital humanities*, that is, when it is a matter of gauging what quantitative methods mean, what they can achieve, and what type of embedding they require. Every type of quantitative analysis, whether implemented for an individual text or a corpus, is based on the operationalization of a clearly formulated question that has to be translated into elementary features, forms, and structures which can be searched, counted, and visualized. With respect to the theoretical inventories presently available to literary studies, it is no coincidence that we can discern a return to structuralism and its sources of inspiration in Russian formalism. Quantitative literary studies was intensely conducted by Russian scientists in the nineteenth century, so that when the formalists made mathematical procedures productive for the analysis of literature, they could already re-

[21] "Every surrogate has its own query potential, which for some purposes may exceed that of the original." (Martin Mueller [see note 3])

[22] This also emphasized by Fotis Jannidis and Gerhard Lauer: "Burrow's Delta and Its Use in German Literary History." In: Matt Erlin, Lynne Tatlock (eds.): *Distant Reading. Topologies of German Culture in the Long Nineteenth Century*, Rochester: Camden House 2014, p. 31.

fer back to highly developed methodological debates.[23] Accordingly, they produced, alongside concrete analyses, theoretical contributions to a critique (in the Kantian sense) of quantitative methods. Boris Tomaševskij, for instance, who applied Markoff's theory of chains (which today are used for, among others, the LDA [*latent Dirichlet allocation*] procedure in topic modeling) to the analysis of prosody, was concerned with a differentiated assessment of the epistemological value of quantitative and qualitative methods:

> Science must not be forbidden the use of any method. ... Number, formula, and graph are symbols of thought just as words are, and they are comprehensible only to those who master the system of symbols. ... Numbers decide nothing, they do not interpret, they are merely a way of establishing and describing facts.[24]

> We must not forget that even in the case of a correct calculation, the figure obtained characterizes merely the frequency of the appearance of the relevant phenomenon, but does not explain anything about its quality.[25]

Treating words, numbers, formulas, and curves as respective symbols of thought with a different epistemological potential corresponds to the understanding of "surrogates" in the concept of *scalable reading* for which additional theoretical impulses can be found in structuralism. Roland Barthes' notion of the "simulacrum" means something systematically entirely similar. The "structuralist activity" is for Barthes "the controlled succession of a certain number of mental operations"[26] through which a text is decomposed into fundamental structural unities whose recomposition then manifests something more than the text form itself does. The structure that is thus worked out through a certain operationalization (in this case: decomposition into particular unities and their modeling) is a surrogate/"simulacrum" of the object, which "makes something appear which remained invisible or, if one prefers, unintelligible in the natural object."[27] In other words, we are dealing with a structural model of the text.[28]

[23] Cf. the excellent depiction by Emmerich Kelih: *Geschichte der Anwendung quantitativer Verfahren in der russischen Sprach- und Literaturwissenschaft*, Hamburg: Kovač, 2008.

[24] Boris Tomaševskij: *O stikhe*. Leningrad 1929, p. 275f. Quoted in: Tzvetan Todorov, "Das Methodenerbe des Formalismus." In: Todorov, *Poetik der Prosa*. Frankfurt/Main: Athenäum 1972, p. 27.

[25] Boris Tomaševskij p. 35f. Quoted in: Tzvetan Todorov [see note 24], *Poetik der Prosa*, p. 27f.

[26] Roland Barthes: "Die strukturalistische Tätigkeit." In: *Kursbuch* 5 (1966): p. 191.

[27] Ibid.

[28] The notion evoked by Barthes in this context that the simulacrum "imitates" the natural object is misleading. Instead, the relationship of each operationalization should be accordingly designated: as formalization, modeling, visualization, and so on.

*Scalable reading* always includes texts and their surrogates, and these surrogates can themselves have the character of a text and thus be read, but it is also a matter of more or less abstract models, formalizations, visualizations or sequences of numbers that must be analyzed and interpreted but cannot be read as literature. Crucial for the success of *scalable reading* as an integrative concept is the question of the degree to which the "scale" can be tapped to measure the epistemological potential of literature in the first place. Is it even meaningful to raise the question of the epistemological potential *of* literature, when it is perhaps more a knowledge about literature that digital methods allow us to acquire in new ways, for instance when Jockers' analysis of metadata corrects the narratives of literary historiography or Schöch's statistical stylometry consolidates the historical knowledge of signals of authorship and genre in challenging ways?[29] Such findings have indeed been at the forefront, which does not fundamentally separate a literary studies that draws upon *digital humanities* from those philological approaches lacking quantitative calculation, but it does separate it from a fundamental conviction of interpretive reading. All deconstructions of hermeneutics aside, the dissemination of literature in schools and universities operates with the basic heuristic formula of the bildungsroman, according to which the critical treatment of texts, which considers one's own perspective alongside others as well as the singular (passages) within the larger context, enables a correspondingly critical relation to the world.[30] By no means must this conviction get lost in a *scalable reading*; many MA programs in German literature that are suffering from declining enrollments are surely happy to be able to list statistical text analysis and data hermeneutics as additional competencies on their website. But in this context it is important that the so-called "structuralist activity" described by Barthes, with its dual operation of decomposition into small units and the knowledge-producing recomposition of something new out of these units, not only allows us to conceive of structural models that could build bridges between qualitative questions and quantitative assessments. Barthes regards structuralist activity as a fundamentally twofold relationship of "creation and reflection."[31] It is production and recepti-

[29] Cf. Matthew Jockers [see note 11], *Macroanalysis*, p. 35-62 as well as Christof Schöch: "Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik." In: Christof Schöch, Lars Schneider (eds.): *Literaturwissenschaft im digitalen Medienwandel, Philologie im Netz, Beiheft* 7, p. 130-157.

[30] Cf. Thomas Weitin: "Die Kunst des Unterscheidens. Kritik und Distinktion in Goethes Wilhelm Meister." In: Zeitschrift für Literaturwissenschaft und Linguistik 166 (2012): 120-149.

[31] Roland Barthes [see note 26]: "Die strukturalistische Tätigkeit," p. 192.

on, writing and reading, that he grasps as the controlled succession of "mental operations" and thus as cognitive processes. In light of the fact that the hardware and software of the cognitive sciences and psycholinguistics have made just as rapid progress as the tools of computational linguistics and the *digital humanities*, these surrogates should also be integrated into *scalable reading*, which introduces an entire spectrum of approaches to the discussion: from cognitive poetics[32] through cognitive hermeneutics[33] to those approaches operating at the crossroads of structuralism and hermeneutics in reception aesthetics, whose potential as a cognitive science seems particularly promising. Wolfgang Iser had originally considered *The Act of Reading* within the purview of psycholinguistics and understood his theory, at precisely the point where he distinguished it from empirical research in the narrow sense, as a suggestion for operationalization.[34] What his theory develops as the "protention" and "retention" that occurs during the reading of narrative texts, as the "wandering viewpoint of the reader" and the function of "empty places" that generate the activity of reading (also thought concretely as a syntagmatic "hiatus in the syntax"[35]), corresponds to a large degree to the experimental designs of observed reading using the *eye tracker*. By measuring fixation (the intake of information), saccades (the gaze changing direction), and regressions (backward saccadic movements), the experiments produce enormous amounts of data that should be confronted with Iser's strong hypotheses.[36] It would then be rewarding for the sake of knowledge about the aesthetics of production to determine the cognitive representation of the implicit reader.

The integration of methods from cognitive science and psycholinguistics into the concept of *scalable reading* provides a further step toward overcoming the unproductive frontline between hermeneutics and statistics. I have to admit that my brief references to points of contact between

[32] Cf. Geert Brone, Jeroen Vandaele (eds.): *Cognitive Poetics. Goals, Gains and Gaps*, Berlin, New York: De Gruyter, 2009.

[33] Cf. Peter Tepe: *Kognitive Hermeneutik. Textinterpretation ist als Erfahrungswissenschaft möglich.* Würzburg: Königshausen & Neumann, 2007.

[34] "As a construction, the theory developed here has not been empirically tested. It is also less a matter of subjecting it to a validity test and more one of helping to design possible grids that would need to be created if one were to conduct empirical studies of the reaction of readers." ["Als Konstruktion ist die hier entwickelte Theorie empirisch nicht überprüft. Es geht ihr auch weniger darum, sich einer experimentellen Geltungsprüfung zu unterziehen, als vielmehr darum, mögliche Raster entwerfen zu helfen, die notwendigerweise erstellt werden müssen, will man empirische Untersuchungen über Leserreaktionen betreiben."] (Wolfgang Iser: *Der Akt des Lesens. Theorie ästhetischer Wirkung.* München: Wilhelm Fink, 1976, p. 9)

[35] Ibid., p. 186.

[36] Cornelius Eggert and Thomas Gilli have done important preparatory work reading Iser and experimenting with the eye tracker in my seminar series on "Hermeneutics – Statistics – Cognition."

empirical research on reading behavior and the theory of the aesthetics of reception adopt the style of those proclamations of scientific prose that at the moment characterizes to a large extent programmatic publications in the direction of *digital humanities*. This is always unsatisfying, yet often unavoidable, because many novices in this field can not yet deliver their own empirical findings and the available results, above all from linguistics, are all too seldom communicated in and with literary studies. Every individual can of course change this, and should make the effort, because the "explicative approach,"[37] which ultimately only imports concepts with recourse to empirical research, does not exhaust the possibilities for a combining of different methods and in the unfavorable case decorates qualitative research with the legitimating rhetoric of empiricism.

## 5. Fast Thinking, Slow Thinking

To conclude my reflections, I would like to discuss findings from cognitive psychology that concern on a fundamental level the relationship of scholarship socialized in qualitative approaches to quantitative methods, because they highlight the psychopathologies of our everyday dealings with statistical knowledge. I am drawing on the analyses of the cognitive psychologist Daniel Kahneman who, after receiving the Nobel Prize for Economics, summarized his findings in the book *Thinking, fast and slow*.[38] Kahneman divides the human brain into two systems, one for calm and observant thought that is responsible for elaborate reflection and consequently for skepticism and self-critique – this is system 2 – and one for intuitive, quickly passed judgments that determine our spontaneous reactions: system 1. For survival and above all for a pleasant life, we of course need both. Problems arise when we activate the quick system 1 at points where the observant thought of system 2 should be at work. And this happens quite frequently, according to Kahneman, whenever we are confronted with statistical facts. Our brain has notorious difficulty dealing with large numbers and their laws. Striving for energy efficiency, it readily hands over the reigns to system 1, which has a penchant for dodging difficult questions and prefers to answer easier ones, all the while suggesting to us that this would be the correct solution. Wherever the law of the large number, statistical knowledge, and logical thought are

[37] Cf. Sophia Wege: "Aufgehender Mond und der Kubikinhalt des Herzens. Zum Verhältnis von Empirie und Literatur in der Kognitiven Literaturwissenschaft." In: Philip Ajouri, Katja Mellmann, Christoph Rauen (eds.): *Empirie in der Literaturwissenschaft*. Münster: Mentis, 2013, p. 395-417.

[38] Cf. Daniel Kahneman: *Schnelles Denken, langsames Denken*. München: Siedler 2011.

called upon, we prefer to cling to small magnitudes, for instance, to actors and narratives that we can best place within a causal order. We thereby quickly find an explanation for a problem.[39] Kahneman makes the tendency towards causal thought, combined with our veritable desire for actors with a specific personal profile that acts according to individual inclinations, responsible for the most common and egregious errors of judgment.

He has conducted a number of similarly designed experiments on this topic, among which the "Linda-Experiment" is the most famous. With his colleague Amos Tverskij, Kahneman gave a test group of students the following description to read: *"Linda is 31-years-old, single, open-minded and very intelligent. She majored in Philosophy. As a student she was very interested in such topics as discrimination and social justice, and she participated in anti-nuclear protests."*[40] The experiment was conducted in the 1980s, at a time when every test subject would have had to think of a student from the University of California in Berkely. And this was exactly the intention of the scientists. They asked their subjects to order a series of possible scenarios for Linda according to their probability. Among them was "Linda is an elementary school teacher," "Linda works in a book store and takes Yoga courses," or "Linda is an insurance agent." The authors of the study had especially isolated two possibilities, namely that "Linda is a teller in a bank" or "Linda is a teller in a bank and active in the feminist movement." The result was as overwhelming as it was devastating for every psychology, statistics, or logic teacher. 89% of the test subjects considered "feminist bank teller" more probable than "bank teller." Although the layout of the questionnaire had afforded them a good opportunity to recognize the simple rule of logic according to which the subset must have less probability than the aggregate, system 2 failed for most of them and they followed the intuition of system 1 to simply cling to the narratively more plausible story. A countercheck with doctoral students in the Decision Sciences program of the Stanford Graduate School of Business yielded almost the same results.

If that is the description of Linda, our intuition would lead us to believe, then she cannot only be a bank teller. That is the narrative of a feminist! And it is indeed. But the most coherent

[39] I am drawing on Albrecht Koschorke's diagnosis, whose general theory of narrative takes as its point of departure the idea that "narrative patterns [. . . ] [function] similarly on the linguistic level to cognitive schema: "Both are techniques of complexity reduction that, to be sure, lead to many mistakes but whose great economic advantage consists in the decrease of time and effort." (Albrecht Koschorke, *Wahrheit und Erfindung*, 2nd edition, Frankfurt am Main: Fischer, 2012, p. 29-30.)

[40] Ibid., p. 195.

story and the most plausible narrative do not necessarily also have the highest probability.[41] In such cases, the inexperienced let themselves be fooled by narrative and ignore the importance of large quantity for their judgment. All feminist bank tellers are part of the entirety of bank tellers. And this large quantity is crucial. Whoever relies on the more obvious law of the small number and judges according to the plausibility of the narrative and its actant, judges falsely. This is called, in the work of Nassim Taleb, a "narrative fallacy."[42]

The example of Linda is innocuous. However, conceivable are many highly important, and hence highly contested, social decisions, about for instance the distribution of resources, technological risk, or those based on statistics of criminality, where our latent inability to deal with large numbers and the cognitive ease with which we rely on narratives and actants are anything but harmless.[43]

This provides clues for how literary studies can treat possible systematic mistakes in the application of two popular theories. The contemporary proliferation of the concepts of "narrative" and "actant" in our research work, the frequent use of these concepts without reflection on their heuristic significance, reveals a need for greater methodological caution. The enthusiasm for our expertise in regard to narrating and narratives is understandable, because these structures are equally literarily, culturally, and socially relevant. And yet the sensitivity for the malfunction of such structures thereby also threatens to disappear, as does the capacity to distinguish between narratives as objects of our research and narratives that are culturally formed (possibly with the participation of literature). In contrast to narratives that lure us to fast thought, literary texts require slow thought. Narrative texts are thus not to be equated with narratives, but are instead in the position to nurture skepticism vis-à-vis narratives.[44] Once we are prepared to get serious technically with the concept of *scalable reading*, statistics will inevitably become a part of our daily business. We do not have to be able to write our own algorithms in order to be able to become competent using a digital tool. But whoever (also)

[41] Cf. ibid., p. 199.

[42] Nassim Nicholas Taleb: *Der Schwarze Schwan. Die Macht höchst unwahrscheinlicher Ereignisse*. München: Hanser, 2008, p. 87-113.

[43] Cf. Paul Slovic: *The Feeling of Risk. New Perspectives on Risk Perception*, New York: Earthscan, 2010.

[44] I have shown elsewhere with reference to the novellas of Conrad Ferdinand Meyer how a form of narration that is skeptical of narrative emerges under the influence of probabilistic reasoning based on statistics (Thomas Weitin: "Verdichtung der Tatsachen. Conrad Ferdinand Meyers Novellenkunst," in: DVjs 89.3 (2015), in print)

analyzes texts statistically, whoever aims to produce and assess data, must be aware of the pitfalls of statistical thought well enough not to get entangled by it. The methods of reading individual texts that are purely oriented towards verification must, if statistics belongs to the "scale" of reading, be expanded to include procedures of critical data analysis, which requires a much higher expenditure of time altogether and a new orientation in the style of thought. Whoever merely seeks to integrate data into our familiar aspirations for verification thinks fast, which can only lead through good luck to lasting results. Whoever thinks slow accepts many fruitless attempts, tests counter-hypotheses, and always only provisionally holds his or her own thesis for reliable. Will this slow thought lead to further advances in literary studies? We shall see.

Transl. by Charlton Payne