Journal of Cultural Analytics

October 25, 2022

From Concepts to Texts and Back: Operationalization as a Core Activity of Digital Humanities

Axel Pichler and Nils Reiter

Axel Pichler University of Stuttgart, axel.pichler@ts.uni-stuttgart.de.

Nils Reiter. University of Cologne, nils.reiter@uni-koeln.de.

Peer-Reviewer:

Dataverse DOI:

ABSTRACT

This article puts operationalization as a research practice and its theoretical consequences into focus. As all sciences as well as humanities areas use concepts to describe their realm of investigation, digital humanities projects are usually faced with the challenge of ‘bridging the gap’ from theoretical concepts (whose meaning(s) depend on a certain theory and which are used to describe expectations, hypothesis and results) to results derived from data. The process of developing methods to bridge this gap is called ‘operationalization’, and it is a common task for any kind of quantitative, formal, or digital analysis. Furthermore, operationalization choices have long-lasting consequences, as they (obviously) influence the results that can be achieved, and, in turn, the possibilities to interpret these results in terms of the original research question. However, even though this process is so important and so common, its theoretical consequences are rarely reflected. Because the concepts that are operationalized cannot be operationalized in isolation, operationalizing is not only an engineering or implementation challenge, but touches on the theoretical core of the research questions we work on, and the fields we work in.

In this article, we first clarify the need to operationalize on selected, representative examples, situate the process within typical DH workflows, and highlight the consequences that operationalization decisions have. We will then argue that operationalization plays such a crucial role for the digital humanities that any kind of theory needs to take off from operationalization practices. Based on these assumptions, we will develop a first scheme of the constraints and necessities of such a theory and reflect their epistemic consequences.

Introduction

All sciences as well as humanities areas use concepts to describe and interpret their realm of investigation. However, while established branches of science and humanities can draw on established practices that determine the relationship between concepts and their instances, digital humanities projects – as projects of a new discipline that combines aspects of computer science with the humanities – are faced with the challenge to establish new ways of ‘bridging the gap’ from theoretical concepts to their instances. This relationship needs to be plausible for all participating disciplines.

The process of developing methods to bridge this gap is called ‘operationalization’ and is the focus of this article. It consists of developing the necessary steps to unambiguously assign the instantiations of a concept to this very concept and thus measure it. Accordingly these steps can be used to detect the instantiations of the concept, allowing its subsequent quantification, manual inspection and/or downstream processing. Such an understanding of operationalization brings together the operationalization-practice from empirical sociology and digital humanities: While, as for example in[1], in sociology operationalization is understood as the development of an instruction for recognizing the observable instantiations of a theoretical concept, in digital humanities, following[2], it is often understood as the development of a rule for measurement (On the history of the notion of operationalization, see:[3]).

The practice of operationalization was first discussed under this name in 1927 by Percy W. Bridgman in his monograph *The Logics of Modern Physics*. With this book Bridgman reacted to what he considered to be the consequences of Einstein’s special and general theory of relativity for the handling of concepts in physics up to that time. In place of the traditional definitional procedure oriented to the central characteristic features of a concept, Bridgman pleaded for a procedure which he exemplified with the help of the concept of length: “The concept of length is therefore fixed when the operations by which length is measured are fixed: that is, the concept of length involves as much as and nothing more than the set of operations by which length is determined. In general, we mean by any concept nothing more than a set of operations; *the concept is synonymous with the corresponding set of operations*”.[4] This definitional practice has far-reaching consequences: it implies a theory of meaning according to which the meaning of a concept/word depends exclusively on the “set of operations” by means of which the instantiations of the concept are recognized and measured. Out of this theory of meaning follows a criterion that might be called the ‘operational criterion of meaning’: According to it, a concept that cannot be transformed into a “set of operations” does not have a meaning at all. We do not follow this strong version of Bridgman’s operationalism, but we see it as complementary to existing concepts of concepts in the humanities. The digital humanities, however, cannot regularly deal with concepts without operationalization.

The terms 'operationalization' and 'quantification' are often used interchangeably,[5] but we want to clearly separate the two in the following: An operationalization is the development of a measurement for (i.e., a procedure to detect) individual instances, such as the classification of a single text, sentence or pair of nodes (see below for a more detailed discussion of examples). Once such a detection has been established, it can be used to collect a large number of instances, which allows quantitative analysis. While this distinction might seem overly pedantic, it does make a difference in some cases, because it is possible to develop a method to quantify a concept in a text, without detecting all individual instances.[6]

Operationalization choices have long-lasting consequences, as they (obviously) influence the results that can be achieved, and, in turn, the possibilities to interpret these results in terms of the original research question. However, even though this process is so important and so common, its theoretical and methodological consequences are rarely reflected.

To do so, one first has to clarify one's understanding of theory and method. The conceptualization of these two concepts is controversial. In the philosophy of science oriented to the natural sciences, however, there is a basic consensus on what is meant by a theory. It reads: "The term 'theory' commonly refers to the form under which scientists express the knowledge resulting from their observations and experimentations in a given domain of phenomena. It is rather uncontroversial that the major function of theories is to allow for the prediction and explanation of the empirical phenomena."[7]

At the latest since the works of Droysen[8] and Dilthey[9], it has been discussed whether such a scientific understanding of theory also applies to the humanities and cultural studies or whether these are confronted with other objects and correspondingly other goals. This debate has been going on for more than a hundred years by now, and we will not delve deeper into it. We believe that it makes sense to follow the minimal consensus on theory outlined above in computational text studies, because of its parallels to data-oriented disciplines such as empirical sociology. In addition, the connection to this understanding of theory offers the advantage that it is accompanied by a certain understanding of method. The latter is understood in this context as a regulated procedure.

Based on these understandings of operationalization, theory and method, we will first introduce three examples of how concepts are actually operationalized (Section 2). The examples are published in scientific/scholarly venues and have been selected to exemplify different forms of operationalization used in current computer-assisted text analysis. While we cannot cover the wide umbrella of text-oriented digital humanities in one article, we believe our selection covers the most common generic approaches. We also want to stress that operationalization in practice can only very rarely be done without making assumptions and taking certain shortcuts, because it is a time-consuming and complex endeavour. Secondly, we will clarify what these forms tell us about the relationship between theory and research practice (Section 3), and, thirdly, develop some recommendations on the evaluation of operationalization practices (Section 4).

This approach is based on two central assumptions: First, we follow the view of authors such as Ted Underwood who proposes that “instead of measuring things, finding patterns, and then finally asking what they mean, we need to start with an interpretive hypothesis (a ‘meaning’ to investigate) and invent a way to test it”[10]. This can be labeled as a ‘top-down-approach’ and is also a core requirement of our previously published article on reflected text analysis.[11] Second, we believe that the approach that is often presented as an alternative, a purely data-driven and inductive approach, a ‘bottom-up-approach’, is a construct: research – even if it is supposedly data-driven – is always based on (at least implicit) theoretical presuppositions. Especially in the digital humanities this becomes obvious: There is no computer program that does not implement theoretical presuppositions, because even the selection of the data set, the selection of a specific computational method and the inspection of some results are the result of decisions that researchers make. Thus, even explorative settings are not purely data-driven, but influenced by (implicit) hypothesis on the usefulness of a certain analysis etc. Operationalization, as we understand it, is the development of a measurement for a given concept. Settings in which existing measurements are used for exploration, no operationalization (in this sense) takes place.

Examples

We will first describe the examples, using information provided in the published articles and – if possible – supplementary material.

Nietzsche's Moral Psychology

The first study whose operationalization practice we want to trace here is Mark Alfano's monograph *Nietzsche's Moral Psychology*.[12] As the title suggests, Alfano aims to reconstruct Nietzsche's moral psychology. To do so, he chooses a method he calls a "synoptic digital humanities approach"[13], which responds to practices in English-language Nietzsche studies that he considers problematic: Following Alfano, research often makes quantifying statements about Nietzsche's use of terms, but these statements are in these cases neither supported empirically, i. e., with exact quantitative data on the supporting passages, nor is it clarified for which concepts Nietzsche used which words. To solve these methodological problems, Alfano develops a three-stage pipeline.

The first stage is the selection of concepts that are paradigmatic for Nietzsche's moral psychology, called "core constructs"[14] by Alfano: "First, I consulted my own previous work and ongoing research for keywords and central constructs. Second, I consulted the secondary literature for further keywords and central constructs. Finally, I shared the merged lists from step one and two with several dozen experts in Nietzsche scholarship and moral psychology, whom I asked for additional constructs."[15] The selection-process thus uses qualitative and scholarly criteria.

In a second step, Alfano then operationalized the selected 'core constructs', "by developing a list of words that Nietzsche characteristically uses to talk about them"[16]. This list was created 'inter-linguistically', which means that he "went through a process of translation and back translation for each core construct and then checked the *Nietzsche Source* for whether Nietzsche used any of the German expressions in question."[17]. Alfano next checks for false positives manually, by inspecting the found passages. In an effort to identify false negatives, he searches for the constructs in English translations and checks wether the passages are included in his collection. A discussion of his translations that do not correspond to Nietzsche's historical state of language is offered by Mattia Riccardi's review of the book.[18].

The actual search via the aforementioned online-edition *Nietzsche Source*[19] is conducted with keyword queries. Technically, Alfano searches for the "words that begin with a given text string by appending an asterisk at the end of the string"[20], thus also searching for potential inflections of the words.

In a third and final step, Alfano determines co-occurrences of the selected 'core constructs'. To do so, he collects all 'core constructs' co-occurring in one of Nietzsche's original textual sections, which are often aphorisms. These co-occurrences are then visualized as a network, with more frequently co-occurring constructs being connected with a stronger edge. This visualization is done for individual books of Nietzsche and for Nietzsche's oeuvre as a whole. On the basis of these data and its visualization, Alfano develops interpretive hypotheses, which he then supports by close readings.

Strangeness Detection

The second project whose operationalization we will reconstruct deals with the literary genre of science fiction texts (SF). In 1972, Suvin[21] coins the term “cognitive estrangement”[22] as a definition for the SF-genre. According to him, SF is a “literary genre whose necessary and sufficient conditions are the presence and interaction of estrangement and cognition”[23]. The term “estrangement” points to the fact that the presented world is not the same as our own, while “cognition” means that, starting from a few assumptions, the world follows rules and natural laws that we can principally comprehend – in contrast to fantasy or mythical stories, which are strange to us, but we also do not assume a coherent set of scientific laws or rules to govern their world.

The article by Simeone et al.[24] starts from this definition and works on an operationalization of narrative “moments”[25], that “feel strange and visionary”[26]. After an unsuccessful attempt using specific technical words, they cast the problem as a supervised, sentence-wise classification task. To this end, they first establish an annotation guideline, which is also included in the accompanying data set.[27] With this guideline, they annotated 138 sentences to measure inter-annotator agreement, and report a Cohen’s $\kappa = 0.63$.[28] The same guideline is then used to annotate about 1500 sentences downloaded from project Gutenberg and the same number of sentences from the Technovelgy collection.

Using these two data sets, they train an SVM classifier,[29] using TF-IDF-scores of the words in the sentences as features, and only the features with the 100 highest scores from each class (which results in 128 features[30] due to overlapping between the classes). Simeone et al. run the experiments with and without the use of a stop word list, and report F$_1$-scores between 58.3 $\%$ and 89.6 $\%$, depending on the used corpus and stop word setting.

Furthermore, they inspect the performance gain provided by each feature, and identify the most important words that indicate strangeness in science fiction. Interestingly, “many of the words [. . .] fall into the category of function words, or words that play a more grammatical role in the sentence and tend to have ambiguous meaning”[31]. Simeone et al. draw the tentative conclusion, that the strangeness in science fiction is not only a matter of meaning, but also of syntax.

Fictionality Detection

As an example for a full-text classification task, we discuss Pipers’ article on fictionality, published 2016 in the *Journal of Cultural Analytics*.[32] The major aim of the article is to find out which textual properties distinguish fictional from non-fictional texts.

The discussion of fictionality as a phenomenon of texts is traced back to Aristotle. This theoretic motivation has no direct impact on the operationalization, however: The actual data that is used for experiments comes from multiple sources. The assignment of texts to the fictionality classes is not discussed in details, we thus assume that these are fiction/non-fiction labels that are provided by publishers and conform to the labels used in the book market. It might be that the operationalization therefore actually aims at the genre distinction fiction/non-fiction instead of fictionality per se.

The core (and main contribution) of the article are experiments to automatically determine wether a text is fictional or not, using a supervised machine learning system (support vector machines, SVM). The 80 lexicon-based features are derived from the commercially available Linguistic Inquiry and Word Count (LIWC) software, which is often used in social sciences. The lexicons contain various linguistic word classes, affect words, and a number of semantic categories like “Family”, “Seeing”, “Religion”, ... The system achieves an accuracy of well above 90 %, leading to the conclusion that differentiating between fiction and non-fiction using lexical material is quite possible (at least in English and German, which was investigated here).

In addition to these raw performance scores, the article investigates which features concretely made contributions to the decision: For each data set, Piper inspects the weight that individual features and feature categories are assigned by the SVM. The most important features that separate fiction and non-fiction in, for instance, the 19th century canon corpus, are features that represent dialogue: Exclamation, question and quotation marks, but also first and second person pronouns.

Comparison and Reflection

After this descriptive summary, we will reconstruct the operationalizations carried out in the selected computer-assisted text analyses and then outline their methodological and theoretical implications. Table 1 summarizes the most important properties of each example, i. e., aspects of the operationalizations. Assuming the understanding of operationalization as introduced in Section 1, the starting point of such an operationalization is one or more (theoretical) concepts which are traced back to phenomena on the text's surface via potentially several intermediate steps. Based on the indicators determined in this way, the operationalized target concept(s) can then be measured. This means that when we speak of measurement here and in the following, we mean only the measurement of the target concept to be operationalized, and neither the measurement of any subordinate concept nor the quantification processes that are involved in many processing steps (such as vectorization). Transforming text data into numerical data can be considered measuring, but it is not the kind of measuring that we mean in this article.

All three examples have in common that their ultimate 'target concept' is theoretically motivated. Alfano explicitly mentions the goal of validating quantitative claims made in Nietzsche scholarship about the "core constructs", Simeone et al. follow a largely hermeneutic account of what science fiction constitutes, and Piper traces the distinction between fiction and non-fiction back to Aristotle. In the way the examples try to achieve these goals they do

Author(s)	Concept	Level	Measurement(s)	Interpretation
Alfano	Terminology	Tokens and pairs of tokens	Lexical rules and co-occurrence	Hermeneutic reconstruction of Nietzsche's moral psychology
Simeone et al.	Strange sentences in science fiction	Sentences	Supervised machine learning	Hermeneutic interpretation of the information gain of individual words
Piper	Fiction/non-fiction	Full text	Supervised machine learning	Data and feature analysis

Table 1: Examples under discussion. The column 'Level' describes the abstraction level of the operationalization, 'Measurement' summarizes the core of the measurement, 'Interpretation' gives information on how the results of the measurement are handled subsequently.

differ. In the first two cases, the operationalization goal is given in an informal way: As a textual description of high-level properties that are only properly accessible for a human expert in the domain. None of the underlying theoretical constructs is provided with an operational definition in the literature: What the concepts under investigation 'mean' can thus only be determined by the set of operations by means of which the instantiations of the concept(s) are recognized or measured. In the last case, the operationalization goal is *motivated* theoretically, but the actual data instances that are fed into the operationalization have an unclear relation to the theory. In many cases, the label fiction/non-fiction is assigned by a publisher or even a library, and given the large size of data set and its heterogeneous origin, a shared theoretical background can almost certainly be excluded.

The instances of the operationalized concepts are to be found on different abstraction levels. The first example combines multiple of these levels: After having identified words or short multi-word expressions as references to "core constructs" in the text, the relation between two "core constructs" is operationalized as a co-occurrence in a (given) textual segment. Thus, two measurements are applied in sequence: First to individual tokens (or multi-word expressions) and second to pairs of previously identified instances of "core constructs". The other two examples are simpler: The second operates on full sentences, and the last one on full texts. But: The reason for using sentences in the second example is a more or less pragmatic one: Conceptually, they are aiming at "narrative moments", and a number of alternative decisions is conceivable (e. g., paragraphs, previously determined scenes[33], a sub set of the sentences, . . .). Even classifying entire text as fictional or non-fictional – as done in Example 3 (fictionality, 2.3) – entails certain limitations: This measurement is not able to deal with non-fictional texts with fictional parts, for instance.

Looking closer at their operationalization workflow, the most striking difference between them is that Alfano does not perform any kind of evaluation with respect to the measurement, while both Simeone et al. and Piper are defining a test and training set and follow best practices established for data science, machine learning and computational linguistics. Apart from that, all the workflows i) start with a theoretical concept and ii) develop or apply ways of identifying its instances in a data set. This results in a measurement, which can iii) either be applied on a new, large data set or can be used to (re-)define the theoretical concepts by means such as feature importance metrics. In either case, the result is then interpreted with respect to the conceptual starting point. Figure 1 shows an idealized abstraction of the conducted workflow graphically. It is important to note that a manual annotation is not a strictly required step in this workflow: If an appropriate data set already exists, as it does in the third example, one can of course re-use it.

Although all three examples conform to this scheme, they differ in the importance of the individual steps. This is especially true for the way each study deals with its central theoretical concepts and how it establishes their instantiations in the data set (which is what we are referring to as 'operationaliza-

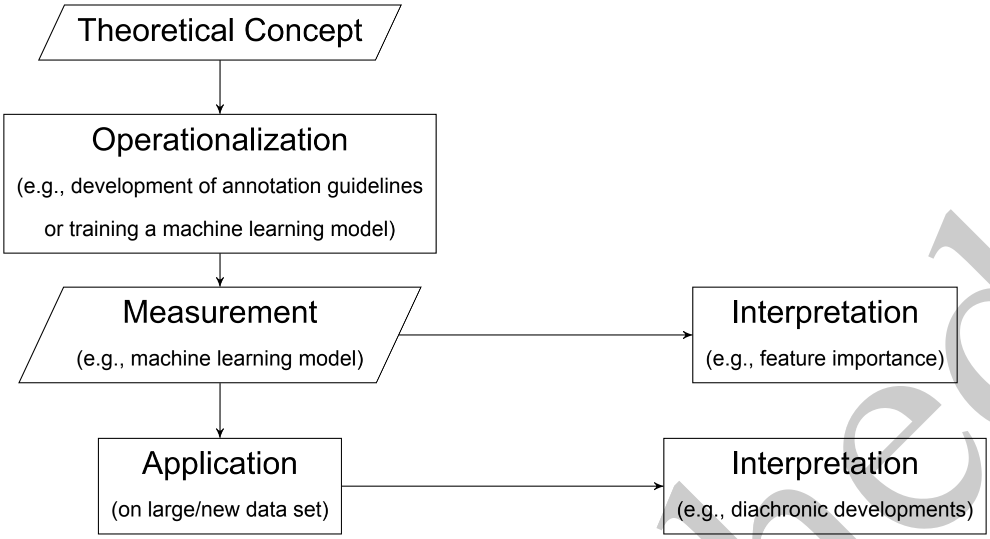
Theoretical Concept
Operationalization
(e.g., development of annotation guidelines or training a machine learning model)
Measurement
(e.g., machine learning model)
Interpretation
(e.g., feature importance)
Application
(on large/new data set)
Interpretation
(e.g., diachronic developments)

Figure 1: Schematic representation of the workflows employed in the discussed examples. Rectangles represent processes, parallelograms data objects (in the widest sense).

tion'). While Simeone et al. develop and apply annotation guidelines for recognizing strange and visionary moments – without defining 'moment', 'strange', and 'visionary' verbally or formally –, the two other studies implicitly presuppose an instantiation of their target concept in different ways: Alfano uses a purely deterministic measurement and assumes that the measurement component themselves (the word lists and rules) encode the concept well enough. Piper assumes that the data set that he uses contains instances of his target concept, such that he can use it as training data for an SVM classifier.

Such a proceeding is consistent with an established and widely used research practice in data sciences. It consists of using pre-existing gold standard data, based on some concept(s), as the training data of a machine learning algorithm, and then using feature importance metrics to explicate the underlying measurement. Following common definitional practices in philosophy, this form of operationalization could be called 'explicative operationalization', because it starts with a vague understanding of a concept that is specified during the course of investigation.[34] It differs from top-down forms of operationalization by the fact that the applied algorithms are understood as established

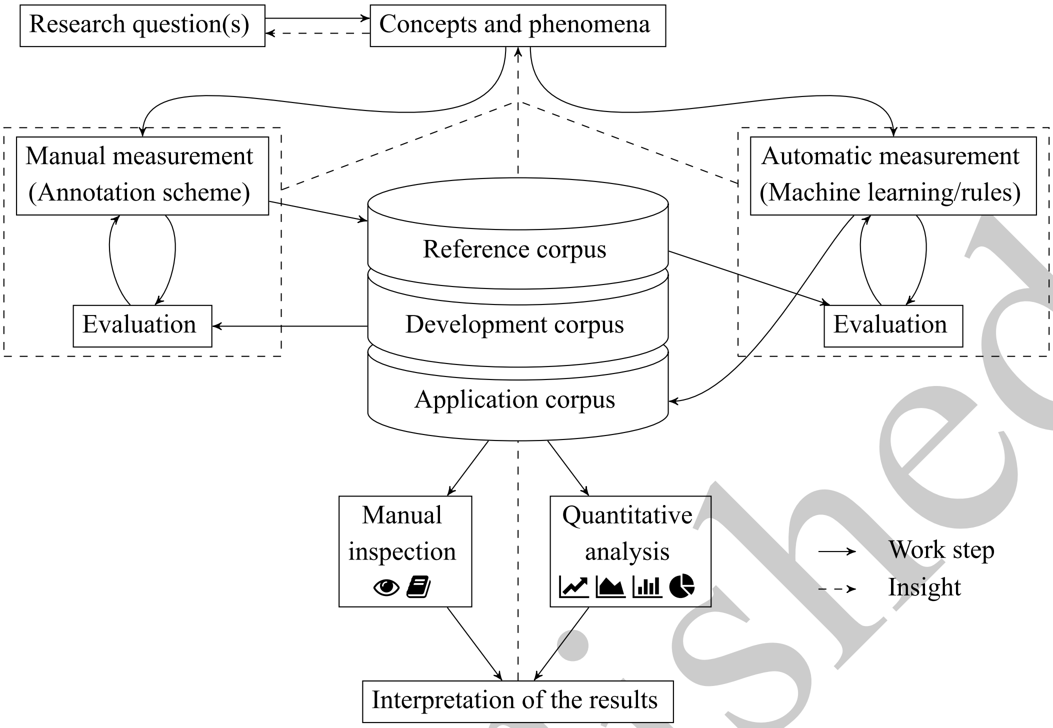

Figure 2: Workflow for reflected text analysis.

tools to specify the concepts of interest[35].

Reflected Text Analytics

After having compared the workflow steps of the examples above, we will abstract from them further by presenting a workflow scheme for 'reflected text analytics'. The scheme is a generic description of such workflows and can give orientation when planning, conducting and comparing operationalizations. The scheme has partially been described in an earlier publication[36] and is depicted visually in Fig. 2.

Starting from a research question, one first identifies the relevant concepts used in the research question. Generally, the workflow we have depicted here allows two paths, and in many cases they are used complementary. As theoretical concepts are abstract and often only vaguely defined, but come with a history of scholarly discussion, it is not straightforward to detect them – manually – in a text. We therefore see the development of rules for measurement or – as we will call it in the following – a 'manual measurement' (on the left of Figure 2) as a first operationalization step, typically done in the form of annotation guidelines, as in Example 2 (strangeness, 2.2). Because the measurement is conducted by humans, the criteria that can be used for detecting and measuring can include things like natural language understanding, world knowledge or even associations readers have. During the operationalization, a data set will be used that we call 'development corpus'. It is meant to be used during the creation of the annotation guidelines, and often consists of canonical and well known texts.

The produced annotation guidelines are then used to establish a 'reference corpus'. This reference corpus can be considered the gold standard (in machine learning terms), and is used to evaluate automatic measurement systems (on the right in Figure 2). It is common practice in the machine learning community to further subdivide this corpus into training and test data or through cross validation. After the automatic measurement has been established, it can be used on the 'application corpus': This is the corpus that is actually of interest with respect to the original research question (e. g., a large corpus of texts from the nineteenth century, as in Example 3 (fictionality, 2.3)). Example 1 (terminology, 2.1), in contrast, directly develops a deterministic automatic detection, without any kind of reference corpus.

It is important to note that these different terms for the corpora describe their *role*: There are cases in which the same corpus is used for all three roles, although the developed measurements are likely to be more generic if different corpora are used.

The application corpus – and the automatically annotated concept instances – are then analyzed further, either qualitatively through inspection, visualization or quantitatively through statistics. Both kinds of findings need to be interpreted before insight is produced with respect to the concepts and – subsequently – research question(s). It is important to note that this interpretation is solely the interpretation of the visualized or analyzed data, and not a ‘proper’ act of literary criticism of a text.

There are several things worth remarking in this schema: i) It is an idealization, and many, if not all, actually conducted text analysis projects deviate from it. This is partially due to the fact that following the scheme to the letter takes several years for any but the most trivial concepts. There are (not marked in the figure) a few obvious shortcuts one can take: If a data set has already been established, it can certainly be re-used (but one should verify that the measured concepts are actually identical). It is also a valid approach to focus solely on manual annotation, and completely ignore the right side. In this case, the established manual measurement would need to be applied to the application corpus. ii) There are two areas in Figure 2 in which operationalization is done: The left cycle to develop a manual measurement and the right cycle to develop an automatic measurement. Both are iterative in nature, such that after a measurement has been established, it is tested and evaluated. After shortcomings have been identified, a new version of the measurement is created. These processes in theory never end. In practice, it is a decision of the involved researchers, if a measurement is good enough for its purpose. iii) Insight and learning does not only happen after the measurement is established and used in unseen data, but also during the operationalization (this is indicated with dashed lines in Figure 2). Example 3 (fictionality, 2.3) can yield as a prominent example for this. The insights produced during the operationalization (e. g., by inspecting errors, fixing bugs or just by being forced to close-read many texts systematically) can even outweigh the results produced from actually measuring. iv) The designation of ‘operationalization’ as a process to establish measuring rules is – despite numerous semantic parallels – not to be confused with rule-based systems known in computer sciences, as measuring rules can be executed by humans and computers alike. It is also quite conceivable that an operationalization results in a set of measuring rules that are (partially) executed by humans and/or by computers.

Evaluating Operationalization(s)

Any operationalization process involves making decisions. Not all of them are directly obvious from the outside, and not all of them stringently follow from the theoretical starting point or are without alternatives. Still, they will have impact on the outcome of an actual measuring. It will thus be of utmost importance for the digital humanities community to compare different ways of operationalizing a concept – and to tell about the theoretical impact of these decisions.[37] For this comparison, we offer the following six criteria to look at. Please note that these are not meant as clear-cut criteria, for which we can determine objectively whether they are fulfilled or not. We see them as discussion points that should be taken into account when evaluating and comparing attempts of operationalization, or when deciding if a measurement is re-used. For the first four aspects, it is generally clear that 'more is better', while the importance of the last two depends on project goals.

Generalizability. A systematic comparison of measurements needs to separate them from the texts (or, more generally, data objects) they are applied to. Thus measurements need to be generalizable from a set of training or development textdata to a larger corpus – within reason: An operationalization of, e. g., strangeness in science fiction should result in a measurement that is applicable to science fiction texts in general (of the same language etc.). The extent to which this is actually the case can in the end only be determined empirically, through validating the measurement on more and untested texts. But during the operationalization, some technical decisions might rule out a generalizability early on. Such an approach to the generalizability of theoretical concepts has numerous parallels to the practice of 'open generalization' propagated by Andrew Piper,[38] with which he follows Geoff Payne's and Malcolm Williams' 'moderate generalization': it consistently reflects the limits of the scope of the operationalized concepts. To give an example: If an LDA model is trained on a corpus, and a specific topic is then identified as being representative of the target concept, this topic is, until further applications have been evaluated, only valid for the training corpus, it does not cover all the semantic dimensions of the target concept.

Example 1 (terminology, 2.1) focuses on the terminology use of a single author (and only develops their methodology in context of Nietzsche's work). There is, however, nothing that prevents us from using the same measurement for other text corpora. We would only need to adapt the initial word lists that represent "core constructs", and potentially find a new way of segmenting the texts, if there is no pre-given segmentation into coherent units. Example 2 (strangeness, 2.2) and Example 3 (fictionality, 2.3) train regular machine learning models, which could be directly applied to new texts, given that a compatible feature extraction can be performed. Obviously, using the models on very different domains or text types will yield reduced performance.

Explicitness. In order to be used by others, including potentially being reimplemented, a measurement needs to be defined exactly and explicitly in all relevant properties. Thus, it needs to be possible to conduct the measurement independently of the person conducting it, with the same results achieved on the same data. The fact that many measurement methods nowadays include randomization (e. g., random initialization of parameters) is not a contradiction to this claim: Non-deterministic systems need to adhere to this in the limit, i. e., repeatedly measuring should yield the same *average* result.

Evaluating the explicitness and in turn the reproducibility is a matter of documentation. It is often the case that not all relevant properties, settings and parameters can be reported within a scientific article, which makes accompanying code and/or appendices important. Ideally, all papers are accompanied by a code repository that contains the entire code needed to reproduce the results in the article. In the case of systems that include supervised machine learning, this also must include the data sets used for training and testing, as is done for Example 2 (strangeness, 2.2). While there might be cases in which legal requirements prohibit the former or the latter, these cases should be kept to a minimum, and potentially use a limited public approach, as Example 3 (fictionality, 2.3) does. Example 1 (terminology, 2.1) presents the full list of used search terms as a table, and is thus explicit enough to be reproducible (which is also important to the author, as he stresses multiple times).

Validity. Validity refers to the link between theory and measurement: "A measuring instrument is considered valid if it measures what it claims it measures"[39]. While looking out for the validity of a measurement is established procedure in the social sciences, this aspect has received much less systematic attention in the digital humanities. This is likely due to the fact that it becomes increasingly clear that high validity in the strictest sense is almost impossible to reach, due to (at least) two reasons: First, many concepts are defined in a highly context-dependent way. In these cases, one can either operationalize parts of the concept, but not all of it, or one specific semantic dimension of a concept that is defined by a certain context. Second, many of the theoretical concepts that we are aiming at are not defined exactly enough in the first place, making it hard to tell what the exact nature of the concept was supposed to be. Still, it is important to maintain clear and defined relations to the theoretical starting points, in order to be able to connect quantitative results to theoretical claims.

In many real examples, the validity of an operationalization is difficult to gauge, because the theory is difficult to grasp. An annotation process by domain experts, as it is done by Example 2 (strangeness, 2.2), is probably the approach that yields the highest validity, because both the intention as well as the extension of a concept are transparently visible. The validity of Example 3 (fictionality, 2.3) is difficult to evaluate, because the theoretical starting point (the concept of fictionality) is not spelled out. One could argue that Example 3 (fictionality, 2.3) is much more an investigation of the labels fiction/non-fiction used by publishers, authors and libraries than of the concept of fictionality per se. The core challenge with respect to validity in Example 1 (terminology, 2.1) is that counting and finding words is not the same as counting and finding concepts. Thus, the link between concepts of interest – "core constructs" – and word lists can be questioned.

Reliability. Measuring the reliability of a measurement refers to its robustness and accuracy: "A research process is reliable when it responds to the same phenomena in the same way regardless of the circumstances of its implementation"[40]. In practice, we measure the reliability through standard evaluation metrics like accuracy, precision and recall, or inter-annotator agreement. Given a large enough corpus, our measurement is confronted with the same phenomena in different circumstances, such that we can determine how it reacts to it. Next to relying on the size of the test data set, different circumstances can also be created artificially, although this is rarely done at this moment. It is clear that high reliability cannot be shown without high explicitness and good documentation.

Example 1 (terminology, 2.1) discussed above does not determine its reliability quantitatively, but given that the whole procedure is deterministic, it should have high reliability. The author is quite aware of both reliability and validity requirements, but argues that domain expertise can compensate for false positives and negatives: "if the researcher is sufficiently familiar with Nietzsche's corpus and observes some uncontroversial safeguards, it should have high validity and reliability."[41] Example 2 (strangeness, 2.2) measures reliability in two ways: They first determine inter-annotator agreement as a measure of the human reliability, and then F1-scores (which are harmonic means between precision and recall) as a reliability measure of the automatic measurement. Example 3 (fictionality, 2.3) re-uses a data set that had been annotated before, but reports the accuracy for the various corpora as a reliability metric for the automatic measurement.

Interpretability. This aspect refers to the fact that some measurements are more transparent than others. A transparent measurement allows its human users to learn why a certain result was determined. In contrast, an intransparent system just gives us a result, without giving these insights. In addition, debugging a measurement is much easier if the scholar/developer can trace its decision making process. This discussion has become more relevant in the most recent past, with the advancements of large neural networks and specifically transformer models. However, it is not an absolute requirement: If a high reliability and validity has been established, a non-interpretable system can certainly be used for large-scale analysis. If the reliability and/or validity of a system is not so certain (e.g., because the data sets are not representative or small, or performance metrics low), we might require more interpretability, as it allows a more robust interpretation of the results.

In addition, the interpretability of a system might be a requirement irrespective of its performance, e. g., for legal reasons. Use cases that do not focus on the prediction on new data sets, but aim at inspecting properties of a trained model are another example. In this case, the interpretability is a hard requirement for an operationalization. All examples discussed above are using interpretable methods. This is most visible in Example 3 (fictionality, 2.3), in which an actual interpretation of the model is conducted. As the same machine learning method is used in Example 2 (strangeness, 2.2), the same level of interpretability can be achieved in principle. But: The interpretability of a support vector machine largely depends on the interpretability of the features used, and using LIWC categories offer much more abstraction than plain *n*-grams. Keyword queries, as are done in Example 1 (terminology, 2.1), are well interpretable, although one has to keep in mind that words without context easily lead to misunderstandings.

Implementability. Finally, we do not consider the automatic measurement to be the only one that requires careful and reflected operationalization. Instead, the manual annotation of the instances of a theoretical concept is just as demanding, requires fine-tuning and experimental development of the measurement. This can easily be seen if the same concept is operationalized for humans competitively, as was done in the shared task SANTA[42]: Eight teams developed annotation guidelines for narrative levels (embedded narratives) independently of each other, and came to very different results – with very different reliabilities. Manual annotation efforts are harder to evaluate, but this does not change the fact that they require great care when conducted. In Example 2 (strangeness, 2.2), a manual annotation is clearly part of the operationalization effort, although it is discussed only as a necessary preparatory step for the automatic measurement. Both Example 1 (terminology, 2.1) and Example 3 (fictionality, 2.3) focus solely on an automatic measurement.

Conclusions

In this paper, we have investigated the role of operationalization for digital humanities projects. We believe that the three examples cover the mainstream approaches to quantitative text analysis. As we argue, operationalization involves a lot of decisions which influence the final result(s). Consequently, we have suggested a vocabulary on how such operationalization decisions can be made more transparent and thus intersubjectively (more) revisable. Finally, we want to summarize the relationship between operationalization and theory in the digital humanities (or at least in the area that is concerned with texts and text analysis).

As we have shown by reconstructing selected examples from the digital humanities, theories or theoretical presuppositions play a central role in these studies in three respects: First, the central concepts that the studies devote themselves to operationalizing or work with are embedded in a larger theoretical context, as the three examples have shown. Second, the studies also draw on theories to the extent that they select them, often explicitly mentioned, as privileged contexts for evaluating and interpreting the results of their data analyses. In addition, third, the applied computer or machine learning models are established by following best practices and standards from statistics, mathematics or computer science, which bring in their own set of assumptions. Irrespective of wether these are actually theories in the sense introduced above, they have consequences that resemble theories and thus need to be reflected in the same way. Accordingly, the theory 'told' by the respective study is the result of the interplay of these theoretical elements. How they actually interact can only be defined by reconstructing the single case of interest. Considering the significance of this interplay such reconstructions would have for the self-reflection and self-understanding of the digital humanities, it is to be hoped that they will soon be realized more frequently.

Notes

[1] Rainer Schnell, Paul B. Hill, and Elke Esser, *Methoden der empirischen Sozialforschung*, 11th (Berlin and Boston: De Gruyter, 2018).

[2] Franco Moretti, *"Operationalizing": or, the function of measurement in modern literary theory*, Pamphlets of the Stanford Literary Lab 6 (Stanford Literary Lab, 2013).

[3] Hasok Chang, *Operationalism*, September 2019.

[4] Percy W. Bridgman, *The Logic of Modern Physics*, 8th (New York: The MacMillan Company, 1958), 5.

[5] A well known example for this equalization is: Moretti, *"Operationalizing": or, the function of measurement in modern literary theory*.

[6] The paper by Federica Bologna, "A Computational Approach to Urban Space in Science Fiction," *Journal of Cultural Analytics* 5, no. 2 (2020), https://doi.org/10.22148/001c.18120 can serve as an example. Bologna quantifies the concept of urban space in science fiction using latent Dirichlet allocation (LDA). She identifies a single topic as the one representing urban space, and looks at the portion of this topic in texts. Thus, she has derived a quantification without being able to pinpoint an exact instance of an urban space in any of the texts.

[7] Marion Vorms, "Theories and Models," in *The Philosophy of Science. A Companion*, ed. Anouk Barberousse et al. (New York: Oxford, 2018), 173.

[8] Johann Gustav Droysen, *Grundriss der Historik*, 1st (Leipzig: Veit und Comp., 1868).

[9] Wilhelm Dilthey, *Einleitung in die Geisteswissenschaften. Versuch einer Grundlegung für das Studium der Gesellschaft und der Geschichte*, 1st (Leipzig: Duncker und Humblot, 1883).

[10] Ted Underwood, *Distant Horizons. Digital Evidence and Literary Change*, 1st (Chicago and London: The University of Chicago Press, 2019), 17.

[11] Axel Pichler and Nils Reiter, "Reflektierte Textanalyse," in *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*, ed. Nils Reiter, Axel Pichler, and Jonas Kuhn (Berlin: De Gruyter, July 2020), 43–60, https://doi.org/10.1515/9783110693973-003.

[12] Mark Alfano, *Nietzsche's Moral Psychology* (Cambridge, UK: Cambridge University Press, 2019).

[13] Alfano, *Nietzsche's Moral Psychology*, 12.

[14] Alfano, *Nietzsche's Moral Psychology*, 14.

[15] Alfano, *Nietzsche's Moral Psychology*, 14.

[16] Alfano, *Nietzsche's Moral Psychology*, 14.

[17] Alfano, *Nietzsche's Moral Psychology*, 17.

[18] Mattia Riccardi, "A Review of Nietzsche's Moral Psychology," (2020), https://ndpr.nd.edu/reviews/nietzsches-moral-psychology/.

[19] Friedrich Nietzsche, "Digitale Kritische Gesamtausgabe Werke und Briefwechsel," ed. Giorgio Colli, Mazzino Montinari, and Paolo D'Iorio, http://www.nietzschesource.org/#eKGWB.

[20] Alfano, *Nietzsche's Moral Psychology*, 16.

[21] Darko Suvin, "On the Poetics of the Science Fiction Genre," *College English* 34, no. 3 (1972): 372–81.

[22] Suvin, "On the Poetics of the Science Fiction Genre," 372.

[23] Suvin, "On the Poetics of the Science Fiction Genre," 375.

[24] Michael Simeone et al., "Towards a Poetics of Strangeness: Experiments in Classifying Language of Technological Novelty," *Journal of Cultural Analytics* 2, no. 1 (2017), https://doi.org/10.22148/16.015.

[25] Simeone et al., "Towards a Poetics of Strangeness: Experiments in Classifying Language of Technological Novelty," 2.

[26] Simeone et al., "Towards a Poetics of Strangeness: Experiments in Classifying Language of Technological Novelty," 2.

[27] The guideline consists of a single paragraph, accompanied by three examples:

> Descriptions or introductions of technology and novel science. Mentions alone of existing tech contemporary to publication will not suffice. We are looking specifically for material or organic inventions in action, making their debut, or explanations for how those inventions work

or set the context for them, not passing mentions of alien life-forms or places that were mentioned before. Description is highly valued. Our approach considers bioengineering. These can exist across sentences. Neologisms that signal the impact of technology and engineering count as perfect past descriptions of their actions.

True: They lifted up, the driver turning the nose of the airjeep in the direction of the flames and explosions and magnesium-lights to the south and tapping his booster-button gently. The vehicle shot forward and came floating in over the scene of the fighting. (Uller Uprising)

False: For two months I had been on the d'Entrecasteaux Islands gathering data for the concluding chapters of my book upon the flora of the volcanic islands of the South Pacific. (Moon Pool)

Close: There was much phosphorescence. Fitfully before the ship and at her sides arose those stranger little swirls of mist that swirl up from the Southern Ocean like breath of sea monsters, whirl for an instant and disappear. (Moon Pool)

[28] Jacob Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement* 20, no. 1 (1960): 37–46.

[29] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," *Machine Learning* 20, no. 3 (1995): 273–97, https://doi.org/10.1007/BF00994018.

[30] This number has been established by the authors of this paper using the code and data provided by Simeone et al.

[31] Simeone et al., "Towards a Poetics of Strangeness: Experiments in Classifying Language of Technological Novelty," 15.

[32] Andrew Piper, "Fictionality," *Journal of Cultural Analytics* 2, no. 2 (2016), https://doi.org/10.22148/16.011.

[33] Albin Zehe et al., "Detecting Scenes in Fiction: A new Segmentation Task," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (Association for Computational Linguistics, April 2021), cf.

[34] The definitional practice of 'explication' is often traced back to the works of Rudolf Carnap's (see for example Rudolf Carnap, *Logical Foundations of Probability*, 1st (Chicago: University of Chicago Press, 1950)). It consists in replacing a vague concept by a clearer and more precise concept that has to fulfill the following requirements: the new concept should be exactly defined, posses certain similarity to the original concept but should be simpler and it should also be fruitful for further investigations.

[35] As we noted in Section 1 digital tools also contain implicit presuppositions, which, as Es, Wieringa, and Schäfer state, should be reflected "in light of, for instance, research activities and reflect […] on how the tool (e.g., its data source, working mechanisms, anticipated use, interface, and embedded assumptions) affects the user, research process and output" (Karin van Es, Maranke Wieringa, and Mirko Tobias Schäfer, "Tool Criticism. From Digital Methods to Digital Methodology," *WS. Proceedings of the 2nd International Conference on Web Studies* 2 (2018): 26). Such an approach has been designated as 'tool criticism'. The workflow of 'Reflected Text Analysis' presented in the next section follows this view, but also emphasizes the need to include in this reflection the theoretical implica-

tions of the respective tools and their relationship to the basic theoretical assumptions of the respective study.

[36] Pichler and Reiter, "Reflektierte Textanalyse."

[37] We have raised this question already in: Axel Pichler and Nils Reiter, "Zur Operationalisierung literaturwissenschaftlicher Begriffe in der algorithmischen Textanalyse. Eine Annäherung über Norbert Altenhofers hermeneutischer Modellinterpretation von Kleists *Das Erdbeben in Chili*" [in German], *Journal of Literary Theory* 15 (2021): 1–29, https://doi.org/10.1515/jlt-2021-2008.

[38] Andrew Piper, *Can we be wrong? The Problem of Textual evidence in a Time of Data*, 1st (Cambridge: University Press, 2020), 55–60.

[39] Klaus Krippendorff, *Content Analysis: An Introduction to its Methodology*, 2nd (Los Angeles, California, USA: Sage, 2004), 313.

[40] Krippendorff, *Content Analysis: An Introduction to its Methodology*, 211.

[41] Alfano, *Nietzsche's Moral Psychology*, 14.

[42] Nils Reiter, Marcus Willand, and Evelyn Gius, "A Shared Task for the Digital Humanities Chapter 1: Introduction to Annotation, Narrative Levels and Shared Tasks," in "A Shared Task for the Digital Humanities," *Cultural Analytics*, November 2019, https://doi.org/10.22148/16.048.