Article

Operationalization and Interpretation Dependence in Computational Literary Studies

Janina Jacke[1]

1. Institute for Modern German Literature and Media, Kiel University, Kiel, Germany.

Abstract. This contribution discusses the relation between the (computational) operationalization of literary studies concepts and the property of interpretation dependence, which applies to many relevant literary studies research questions and poses specific challenges to operationalization. Using the attempt to operationalize the narratological concept of unreliable narration as an illustrative example, the paper addresses three crucial subtasks for operationalizing a concept (definition, identification of steps necessary to decide if it applies and the actual application) and explicates if and why interpretation dependence complicates them. The paper concludes with general recommendations for operationalizing complex concepts coupled with a high degree of interpretation dependence.

1. Introduction

The question of how computational literary studies relates (or should relate) to 'traditional' literary studies' interests, theories and methods is a subject of constant debate (cf. e.g. Trilcke and Fischer 2016, 4–8 for an overview). This article is based on the assumptions that computational literary studies can and should in fact relate to core research interests of 'traditional' literary studies, and that it is worth going to some lengths to argue and demonstrate why and how this is possible. Since the interpretation of literary texts is one of the central concerns of literary studies and often poses particular challenges for computational approaches, this article discusses the relationship between the (computational) operationalization of literary studies concepts[1] for text analysis/-classification and the property of interpretation dependence. Operationalization is defined here as the threefold task of developing an adequate intensional definition (i.e. a definition specifying the sense of the concept, usually by providing necessary and sufficient conditions, Cook 2009, 155), identifying the steps necessary to decide whether a concept applies to a text or text passage and carrying out this decision on a text or text corpus. Interpretation dependence is understood as a gradable property of statements about literary texts or of literary studies concepts. Its degree is defined by the extent to which the attribution of a concept to a text or a statement about a text depends (in terms of justification) on non-truth-preserving inference or (controversial) contextual assumptions. The main aim of the paper is to explain how (high-degree) interpretation

---

1. "Concept" is understood here as an abstract idea that is "crucial to such psychological processes as categorization, inference, memory, learning, and decision-making" (Margolis and Laurence 2023). I follow the classical theory of concepts, assuming that a concept "has definitional structure in that it is composed of simpler concepts that express necessary and sufficient conditions for falling under" it (Margolis and Laurence 2023).

dependence creates challenges for the operationalization of literary studies concepts and to suggest possible ways of dealing with these challenges.

The article is structured as follows: As the notions of operationalization and interpretation dependence are central to this article, I will briefly sketch the debates concerning these concepts in the fields of Digital Humanities/CLS resp. literary studies/analytical philosophy of literature and explain how my own definitions relate to previous definition attempts (subsection 2.1 and subsection 2.2). Based on these explanations, I will comment on the method I'm following in this paper, which combines theoretical ('top-down') considerations with pratice-based ('bottom-up') elements in a way that will need to be explained in more detail (subsection 2.3). The practice-based elements originate from the ongoing project to operationalize the narratological concept of unreliable narration (Jacke 2023a). Roughly speaking, this concept is meant to denote fictional narratives in which the narrator is not to be trusted, and it is often described as a highly interpretation-dependent concept (Kindt 2008).

The following sections are based on the different tasks that (according to my definition of "operationalization") need to be carried out when operationalizing a concept: I will first discuss the task of developing an adequate definition (section 3), including an explanation of why this task might be perceived as being related to interpretation dependence, while I am arguing that it is, in fact, not. The next section will be devoted to the task of identifying the steps necessary to determine whether a concept applies to a literary text (section 4). In describing this task, I will also explain and illustrate the notion of interpretation dependence in more detail and show why and how a high degree of interpretation dependence significantly complicates this task. The next part of the paper focuses on the step of applying a concept to a text (corpus) – and sketches out ideas for dealing with potential problems (section 5). The paper concludes with a summary and an attempt to formulate general recommendations for operationalizing complex concepts coupled with a high degree of interpretation dependence (section 6).

2. The Central Concepts and Methods of this Paper

Before going through the different tasks that are necessary to operationalize a concept and analyzing how they are affected by the concept's interpretation dependence, I would like to comment on and contextualize the concepts and methods that are central to my approach, starting with the notion of operationalization.

2.1 Operationalization

As Pichler and Reiter (2021, 4) have pointed out, operationalization is a concept that is often mentioned in recent DH and CLS papers but rarely specified or problematized. The concept was introduced to the field of DH by Moretti in his paper *'Operationalizing': or, the function of measurement in modern literary theory* in 2013. According to Moretti, "[o]perationalizing means building a bridge from concepts to measurement, and then to the world. In our case: from the concepts of literary theory, through some form of quantification, to literary texts" (Moretti 2013, 1). In other words: Operationalization (in CLS) is the attempt to take literary studies concepts as a starting point and translate them in a way that they can be computationally measured, or quantified, in texts. Put this way,

operationalization may at first glance sound like a promising way for CLS to maintain the connection to literary studies and yield results with the potential of being perceived as relevant from a more traditional point of view. However, CLS researchers have since shown that neither is Moretti's concept of operationalization well-considered enough, nor is his own attempt to operationalize the literary studies concept of character-space convincing from a literary studies perspective. Let's look at these two observations in more detail, starting with the conceptual shortcomings.

Moretti himself mentions that the concept of operationalization has its origin in Bridgman's monograph *The Logic of Modern Physics* from 1927 (Bridgman 1954). As Krautter (2022, 222–225) carves out, however, Bridgman's own initial concept of operationalization was a very strong and not thoroughly plausible one: According to Bridgman (1954, 5), any (physical) concept should be defined by providing the (one) set of operations that need to be carried out in order to find out if the concept applies resp. in order to measure the concept. Bridgman thus demands solely operational definitions for physical concepts, deviating from the classical account of (intensional) definitions that provide the necessary and (together) sufficient conditions (i.e. features) an object needs to have to fall under a concept. Now as Krautter (2022, 225) argues, it is very difficult to verify whether the operations provided in the context of an operational definition are actually adequate to measure a concept if we don't have a definition of the concept that is independent of these operations. While Bridgman later softens his approach (Krautter 2022, 225–226), Moretti doesn't comment on these problems and doesn't clarify his own position in this context.

This leads us to the second problem that has been attested to Moretti: Moretti does not specify the relation between the original concept and his operationalization. It appears that he doesn't in fact operationalize the concept of character space but only measures something that may show some (unspecified) relation to it but underwent a significant reduction in complexity, for example centrality (cf. e.g. Trilcke and Fischer 2016, 3, 11–16; Krautter 2022, 228–234), and he doesn't discuss any methods to evaluate the success of his operationalization attempt (Krautter 2022, 232).

Examples like this are very counterproductive if the goal is to build a bridge between traditional literary studies approaches and CLS, as unreflected reduction of complexity is a common concern voiced e.g. by traditional literary scholars in connection with CLS (Gius and Jacke 2022). Krautter argues that, in order to avoid this problem, we need a notion of operationalization that is neither too restrictive (as is Bridgman's initial account) not too permissive (as Krautter assesses Gius's 2019 account). Krautter thus seconds the notion of "operationalization" that Pichler and Reiter offer in their 2021 paper: That operationalization is "the development of procedures that, explicitly and rule-based, trace a term back to text surface phenomena, potentially by breaking it down into several sub-steps or -terms" (Pichler and Reiter 2021, 4, my translation).[2]

I would like to argue, however, that the focus on text surface phenomena may make this definition too narrow. It has often been pointed out that many literary studies concepts are notoriously complex, which makes their operationalization extremely *difficult*, but if it were a necessary condition for an operationalization that a concept can be fully traced

2. Original: "die Entwicklung von Verfahren, die einen Begriff über potentiell mehrere Teilschritte oder -begriffe explizit und regelgeleitet auf Textoberflächenphänomene zurückführen".

back to text surface phenomena, this would make a complete operationalization of most concepts impossible. Pichler and Reiter (2021) themselves mention the difference of operationalizing a concept for humans vs. for computers. While computers may depend fully on text surface phenomena to decide whether a concept applies to a text or text passage, humans don't. So an operationalization for humans does not have to trace a concept back to text surface features only, but may also include sub-steps or -concepts that are based on extratextual elements or prior knowledge. In a later paper, Pichler and Reiter modify their definition of "operationalization", forgoing the reference to text surface phenomena. They define "operationalization" as "developing the necessary steps to unambiguously assign the instantiations of a concept to this very concept and thus measure it" (Pichler and Reiter 2022). This definition, however, contains a new element that may make it too narrow again. I would like to suggest that the notion of operationalization for the field of (C)LS should not contain as a necessary condition that the instantiations of a concept should be assigned *unambiguously*. As I will argue in detail later, many literary texts are (partly) ambiguous or indeterminate. This may often lead to cases in which even the most thorough operationalization attempt will not provide ready-made unambiguous answers on whether a concept applies to a text or text passage.

Against this background, I would like to repeat and briefly comment on the definition of "operationalization" that I am proposing in this paper: Operationalization is the threefold task of developing an adequate intensional definition, identifying the steps necessary to decide whether a concept applies to a text or text passage and carrying out this decision on a text or text corpus. My own suggested definition of "operationalization" of literary studies concepts does without the above-mentioned elements that may make it too narrow: It allows for the possibilities that an operationalization may need to include references to extratextual elements and that it does not always enable an unambiguous identification of a concept's instantiations. In addition, my definition highlights the need to develop an adequate non-operational definition of the concept as a first step, in order to provide a reference for evaluating the operationalization's accuracy of fit. Finally, by explictly including the step of actually carrying out the decision of whether a concept applies to a text or text passage, my definition not only includes a proof of concept in terms of the concept's applicability. As I will argue in more detail in subsection 2.3, section 4 and section 5, such an application also introduces a (first) feedback loop to identify necessary steps that could not be identified top-down, i.e. solely based on theoretical reflection. Based on the assumption that the common core of previously suggested definitions of "operationalization" is providing the 'application conditions' of a concept, my definition can thus be classified as an explication (Carnap 1965, paragraph 2) in that it aims at capturing the relevant aspects of the term's previous use while making it more precise.

After having contextualized my definition of "operationalization", I will now turn to the second central notion in this paper: the notion of interpretation dependence.

2.2 Interpretation Dependence

As I have spelled out in the introduction, the aim of this paper is to shown how the operationalization of a literary studies concept is complicated by the interpretation

dependence of the concept or its application (and to suggest ways of dealing with these complications). It will thus be necessary to say a few words about the notion of interpretation/interpretation dependence in literary studies.

The interpretation of literary texts is regarded as one of the core activities of literary studies (cf. e.g. Weimar 2007, 486). While it has often been pointed out that "interpretation" (in the context of literary studies) is an ambiguous term (cf. e.g. Bühler 2003), Descher et al. (2015, 23–24) suggest that all forms of interpretation include the attribution of meaning. One very basic differentiation is the one between interpretation as a process of attributing meaning vs. as the result of such a process (Spree 2007, 168). Usually, not all kinds of attribution of meaning are regarded as interpretation. For example, Spree (2007, 168–169) suggests as a necessary condition of his concept of interpretation that it is aimed at the text as a whole and at non-obvious aspects of meaning that exceed its lexical word sense. Especially the latter criterion can be seen as an attempt to distinguish the interpretation of literary texts from their description or descriptive analysis (cf. Kindt and Müller 2003 for an overview).

It is my impression that one central intuition that the debate surrounding the distinction between text description and interpretation attempts to capture is that descriptions can easily be agreed upon (as they can be verified or falsified relatively easily) while interpretations are debatable and not always easy to agree on. This difference is also crucial in the context of CLS: Computers can handle many descriptive tasks comparably well but usually aren't expected to be able to find answers to questions about texts that even humans can't agree on. This is also the reason why inter-annotator agreement for a task is usually measured to assess a computational model's results on the same task, or why gold standard annotations are used to train machine learning models.

An account that, in my view, grasps this relevant difference between description and interpretation very well has been offered by Reichert (1969). In close examination of an argument by Weitz (1964), Reichert (1969, 184) suggests that descriptions are logically independent of explanation or can be supported by referring to the words of the literary text, while "interpretive propositions [...] are logically related to other propositions about the work [...]. And the plausibility or force of explanatory statements depends in part upon the validity of the logic and the plausibility of the other propositions themselves" (Reichert 1969, 282). He adds that, usually, "interpretations [...] depend for their validity upon the truth of the more general hypotheses (historical, psychological, or aesthetic, for example) from which they are derived. These general hypotheses being themselves uncertain, the interpretations that critics derive from them are equally, if not more uncertain" (Reichert 1969, 286). In other words, Reichert's distinguishing criterion is that to argue for the truth or plausibility of an interpretive statement, we usually need to provide premises that lie outside the text itself and that are often debatable themselves. It is important to note that Reichert does not talk about how the interpretive statements are being developed (i.e. the process) but about how these statements (i.e. the results of a process) can be justified (on the difference between discovering and justifying hypotheses in DH, cf. Gerstorfer 2020). Also, Reichert emphasizes that "we ask for interpretations and explanations of individual words, lines, and speeches" (Reichert 1969, 285), so in contrast to Spree's account of interpretation, Reichert's is not limited to statements about literary texts as a whole. Before I expound my own

definition of interpretation (or rather: interpretation dependence), I want to highlight one final aspect of Reichert's account that informs my own: Reichert talks of a "pyramid of statements [about literary texts, J.J.], building on the solid foundation of the words and toward the more and more patently interpretive" (Reichert 1969, 285). This suggests that description and interpretation are not a dichotomy but poles of a scale and thus statements about texts can be more or less interpretive (or: interpretation-dependent).

Against this background, let me iterate and explain my own definition of interpretation dependence. I will not flesh out all the details here, but focus on relating it to the debate outlined above – an in-depth explanation will follow in section 4 below. My suggestion is that interpretation dependence be understood as a gradable property of statements about literary texts or of literary studies concepts. Its degree is defined by the extent to which the attribution of a concept to a text or a statement about a text depends (in terms of justification) on non-truth-preserving inference or (controversial) contextual assumptions. I will comment step by step on my definitional decisions. First of all, I decided to introduce the concept of interpretation *dependence* instead of interpretation to highlight that my concept deviates in some regards from Spree's standard definition. My concept is not meant to substitute narrower concepts of interpretation like Spree's but is supposed to incorporate different narrower concepts and provide a way to describe and explain some of the relevant features of their instantiations (i.e. of concrete interpretations) in more detail. Second, following Reichert, I opted to define interpretation dependence as a gradable feature. Not only does this conform to the intuition that statements about texts can be more or less debatable, more or less interpretation-dependent. As Carnap (1959, 16) highlights, comparative and quantitative concepts also enable a much more precise description of complex phenomena than classificatory concepts do. Third, even though Descher et al. (2015) see the reference to the meaning of a literary text as the potential common core of interpretation concepts in literary studies, I opted against including it in my definition of interpretation dependence. One reason is that I modeled interpretation dependence as a gradable feature, while the question of whether a statement is about the meaning of a literary text seems to require a yes-or-no answer. The second reason is that the concept of meaning itself is non-trivial and would require further clarification, which I will not dive deeply into in this paper. I will, however, talk a bit more about how two types of meaning (related to content-specifying and content-transcending interpretation) influence the degree of interpretation dependence. Fourth, I have so far only talked about statements about literary texts. In my definition, I am adding literary studies concepts as possible bearer of interpretation dependence. This is important in the context of this paper, as concepts are what is operationalized in CLS. The relation between literary studies concepts and statements about literary texts is that statements can be about whether a concept applies to a text or text passage. As I will argue in section 4, if a concept is interpretation-dependent, the statement in which it is used will also be. Fifth, my definition is not restricted to statements about the literary text as a whole but allows for interpretation-dependent statements to concern parts of texts. Sixth, my definition follows Reichert's suggestion that the property of interpretation dependence is directly related to the more complex and debatable argumentative foundation of certain statements about texts. Reichert mainly mentions extratextual (and potentially controversial) contexts that serve as premises. What I am adding is a second element that makes the argumentative foundation more debatable,

namely non-truth-preserving inferences that may be necessary to move from premises to the conclusion. Both elements (i.e. controversial extratextual premises and non-truth-preserving inference) offer great potential to push the concept of interpretation dependence more in the direction of a quantitative concept. I will elaborate on the model (including the notion of non-truth-preserving inference) in more detail in section 4.

While, as I argued above, my definition of "operationalization" is very clearly a Carnapian explication, the case is not equally clear for my definition of "interpretation dependence". While my concept aims to grasp relevant intuitions concerning the difference between the description and interpretation of literary texts, it can be argued that I am introducing a new term that highlights the 'gradable feature' quality that I am aiming at.[3]

After having contextualized and argued for my definitions of "operationalization" and "interpretation dependence", I will conclude my preparatory remarks by saying a few more words about my intentions in this paper, my underlying assumptions and the methods I chose to implement them. This will make it easier to assess the validity/plausibility of my argument and theses.

2.3 Assumptions, Aims and the Interplay of Theory and Practice in this Paper

As I indicated in the introduction, my main aims of this paper are (1) to show how the feature of interpretation dependence that literary studies concepts or statements about literary texts can exhibit complicates the task of operationalizing such concepts, and (2) to sketch out suggestions about how we can deal with these kinds of problems. These aims are based on the assumption that it is possible to bridge the gap between traditional and computational literary studies and on the normative position that we should attempt to in fact bridge this gap. A further assumption is that one way to contribute to bridging the gap is to show how CLS can contribute to interpretation-related actions in literary studies. To do this, I need to argue that it is not self-evident that CLS can contribute to interpretation-related questions (e.g. because interpretation dependence complicates the operationalization of concepts), but that it is possible to fruitfully address these problems. Two secondary aims that are preconditions for being able to work on the main aims are to provide definitions of "operationalization" and "interpretation dependence" and to argue for them, which I have done in the two previous subsections.

I will now take a closer look on my main aims to explain how theory and practice interact in this paper. The first main aim is *theoretical* in nature: I want to provide an analysis of the interdependency of different phenomena at a significant level of abstraction. I want to show how, for conceptual reasons, operationalization is affected by interpretation dependence. However, (computational) literary studies *practice* comes into play in different ways. First, due to the level of abstraction that comes with theoretical thoughts, I will illustrate my ideas by referring to a specific example concept and the attempt to operationalize it, namely unreliable narration. However, the plausibility of

---

3. Gius (2016, 13) uses the term "interpretation dependency" to describe a feature of narratological annotation categories, but she does not define the concept explicitly. In her approach, interpretation dependency values between 1 and 5 are assigned to narratological annotation categories "according to the respective insights from their application".

my theoretical argument concerning the general relation between operationalization and interpretation dependence (and the normative component suggesting general ways of handling occurring problems) does not necessarily hinge on whether the operationalization of unreliable narration that is presented here is deemed convincing: In its illustrative function, the example of operationalizing unreliable narration is meant to merely provide a sense of how operationalization problems (related to interpretation dependence) may look in practice. Second, the theoretical theses I am presenting in this paper are in their genesis informed by my previous experience in working on and with literary studies concepts and interpreting literary texts, also in connection with unreliable narration. However, the justification of theses does not hinge on their way of discovery. To explain the third way in which practice comes into play in this paper, I have to anticipate parts of the theses I develop in implementing my main aims. As I am spelling out in more detail later, the first task of operationalizing a concept (finding a definition) can mostly be carried out in a top-down manner, i.e. based on abstract and conceptual thinking (even though it may, and often will, include practice-oriented considerations, e.g. the previous use of the relevant term in literary studies practice or the fruitfulness of the defined concept when practically applied). For the second task (identifying the steps necessary to decide whether a concept applies), it is usually possible to start in a top-down manner again, working with what follows from the definition of the concept. However, especially when we move from the question of what we need to know to decide whether a concept applies to the question of how we can find these things out (see section 4), top-down reasoning will often not be enough. As I will show and argue later, it is often necessary to observe, analyze and (sometimes) dispute and re-shape the actual practice of applying a concept in literary studies. This will not only provide further insights into the question of how different types of specific qualities in literary text have an impact on the steps necessary for deciding whether a concept applies. It can also help identify (types of) assumptions and forms of inference that are used in deciding whether a concept applies. This is one reason why I included the task of carrying out the decision of whether a concept applies as the third and final task of operationalization: It provides a test to see if the identified steps were precise enough, and if they weren't, it can help to further specify them. For this reason, a 'switch of modes' may be perceived between section 4 and section 5 of this paper, because I switch from mainly theoretical, top-down reasoning to a relatively detailed description of the experimental three-tracked approach of operationalizing unreliable narration in the project CAUTION (short for Computer-aided Analysis of Unreliability and Truth in Fiction – Interconnecting and Operationalizing Narratology) that serves as an illustration of how bottom-up reasoning may come into play here. At the time this article was written, the operationalization of unreliable narration in CAUTION has not yet been concluded, so the success/plausibility of this specific operationalization itself cannot be fully assessed yet. I believe, however, that, again, the plausibility of my theoretical argument (here: that the task of carrying out the decision of whether a concept applies can assist in refining the necessary steps) does not fully hinge on the plausibility of this specific (ongoing) operationalization attempt.

After having clarified and contextualized the relevant concepts and methods of this paper, I will now go through the different steps that are, according to my notion of operationalization, necessary to operationalize a concept and show if and, if so, how

interpretation dependence comes into play.

3. Finding a Definition

The first task of operationalizing a concept, i.e. finding an adequate intensional definition (usually by providing the necessary and sufficient features an object has to have to be an instantiation of the concept), is not always an easy one to begin with, as many concepts in literary studies lack explicit definitions. Typically, such concepts are only roughly described, with different features and variants of the concept being unsystematically mentioned and used in articles or books. It is often not made clear whether the characteristics mentioned are necessary or sufficient conditions[4] of the concept, or whether they are, e.g., merely typical features. This problem is aggravated by the fact that concepts are often characterized merely by giving example cases, and it is not made explicit which features of the given case make it a part of the extension of the concept.[5] One reason for this lack of explicit definitions in many literary studies publications may be that literary studies tends to approximate its meta-language to its object language (Fricke 1970), i.e. often attempts to make research publications aesthetically pleasing – and elements that are perceived as too formalistic (such as definitions) may undermine this aspiration.

Further confusion is caused by the fact that literary studies terminology is not necessarily used in a standardized way, meaning that the same term may have different definitions (and sometimes: different concepts, i.e. their extensions differ). This may be less of an issue with terminological neologisms (such as "homodiegetic", Genette 2010, 158–159), but is very much an issue with literary studies terminology borrowed from natural/everyday language (such as "unreliable", see below).

The task of developing an explicit definition (i.e. providing the defining features) for a concept is characterized by the challenge of balancing different criteria: Since we are usually dealing with pre-existing terms/concepts, we need to ensure that our definition is sufficiently similar to previous uses of the concept while at the same time avoiding some of the formal problems of previous definitions, such as lack of clarity. What we need are, basically, Carnapian explications (Carnap 1965, paragraph 2; Gerstorfer and Gius 2025) – but both the interpretation and weighting of his quality criteria (exactness, similarity, fruitfulness and simplicity) may be different (and especially challenging) for literary studies and other humanities disciplines (Jacke 2019, 290–298). In the context of the (computational) operationalization of a literary studies concept, it is particularly important to keep in mind that a definition should respect the relevant intentions and goals of literary studies in relation to the concept, and not maximize the chances of high inter-annotator agreement or automated recognition at their expense. Let me briefly elaborate on this assumption: While Carnapian explications usually generate concepts that deviate from previous uses to some degree, e.g. for the benefit of increased precision, neglecting the original literary studies intentions behind a concept will risk further estranging of traditional and computational literary studies. The key point

---

4. A feature is necessary for a concept to apply if it cannot apply without that particular feature. A set of features is sufficient for a concept to apply if that particular set of features occurs *only* when the concept applies. Specifying the necessary and sufficient conditions for a concept leads to a classical definition (Brennan 2022).
5. The extension of a term/concept is the sum of the entities to which the concept applies.

here seems to be that, in the field of literary studies, the fruitfulness criterion is very heavily weighed: Literary studies concepts should help us to identify and talk about relevant, interesting (textual) phenomena, and those phenomena are often complex. To sacrifice fruitfulness for exactness thus goes against a crucial aspect of literary studies disciplinary self-conception. This is even more true if exactness comes in the form of maximized agreement, as textual ambiguity and a resulting pluralism of interpretation are also valued highly in literary studies. So if the maxim that literary studies intentions behind a concept should not be neglected in favor of maximized agreement is refuted, it should be for very good and well-argued reasons.

One way of dealing with this problem, i.e. of mediating between complexity/fruitfulness and exactness, is to use very inclusive definitions (i.e. definitions with a large extension) and/or disjunctive definitions (i.e. definitions that are organized as alternatives of subtypes if no common necessary and sufficient conditions for the general concept can be found).[6] It is often advisable to select single types or sub-concepts of complex concepts for operationalization and make explicit how they relate to the complex concept as a whole or to other subtypes.

For the concept of unreliable narration, an inclusive and complex definition by weighing the quality criteria for explications in the context of literary studies has been proposed (Jacke 2019, 289–308). Since the resulting definition reveals the heterogeneity of the concept (we might actually be dealing with different concepts subsumed under one term after all), it has been decided to single out one subtype of unreliable narration for the attempt of the (computational) operationalization in the context of the CAUTION project (Jacke 2023a).[7] This subtype can be called *incorrect assertion*, and it applies to a sentence of a fictional text if and only if in that sentence the narrator asserts a proposition about the fictive world of the text that is incorrect in that world.[8] This excludes from the study, for example, cases in which narrators (merely) omit relevant elements of the fictive world, (merely) hold incorrect beliefs about the fictive world, as well as cases in which the narrators' expressed (through words or actions) or internalized values are in conflict with the text's message – all of which are often referred to as "unreliable narration". Defined in this way, this (sub-)concept is suitable for use in the context of text analysis, as it can help to categorize sentences or text passages. To use it for text or narrator classification would require the definition of a grading system and/or threshold value to transform the analysis results into a synthesized label for a text or narrator.

It should be noted that the development of a definition might always be perceived as a reduction in that it does exclude some previous uses of the term in question (even in the case of the most comprehensive definitions that pay a lot of attention to previous uses

---

6. This type of definition has some similarities to a taxonomy – but the criteria of exclusivity and completeness will typically not be met at every level.

7. The project is carried out in collaboration with Jonas Kuhn (computational linguistics, University of Stuttgart).

8. The questions of what it means for an assertion to be true and what truth means in connection with fictional utterances spark complex debates in the fields of epistemology and philosophy of language. In debates concerning the notion of truth in fiction, discussed questions are if and in which regard fictive entities exist and if (and, if so, in which regard) assertions about such entities can be true (cf. e.g. Lewis 1978; for an overview Kroon and Voltolini 2023). I will not delve into these debates here. It is my impression that we share an intuitive understanding of what it means that the utterance of a narrator is true in a fictive world of a narration: We imagine that the assertions of the narrator are about the details of a fictive world – and we can ask whether the narration reports this world correctly.

and intentions). This careful reduction is a necessary step to ensure a fruitful use of terminology in academia. In the context of our project, we also have decided – among other reasons due to the heterogeneity of the concept of unreliable narration – to choose a sub-concept for operationalization. This may be a reduction on a *practical* level (as we are not trying to operationalize unreliable narration in its entirety), but not on a *conceptual* level (we are not suggesting that only false claims should count as unreliable narration.) Thus what can at best result from the project CAUTION can only be a partial operationalization of unreliable narration, though maybe a complete operationalization of the subconcept incorrect assertion.[9]

Now, the challenges outlined in this section might be seen as resulting from the interpretation dependence of the concept in question. For example, if two literary scholars disagree about whether a particular text passage is a case of unreliable narration, the reason for this may well be that they have different definitions of the concept in mind (Gius and Jacke 2017, 246–250) – yet such disagreement is often taken as an indicator or even proof of interpretation dependence. For the sake of clarity, however, I propose to distinguish between unclear concept definitions and interpretation dependence,[10] the latter of which I will discuss in more detail in the following section.

4. Identifying the Steps – with a Special Focus on Interpretation Dependence

Having found or developed an appropriate definition and, if necessary, chosen from a variety of sub-concepts, the second task of operationalization will be to identify the various steps that are necessary to determine whether the concept applies. In the case of literary studies, the relevant questions are (1) what we need to know in order to decide whether a text or text passage exemplifies a concept and (2) how we can find out these things.

The first question can usually be answered if we have a good definition of the concept at hand. Take, for example, the definition of *incorrect assertion*: A sentence in a fictional text is a case of the unreliable narration subtype *incorrect assertion* if, and only if, in that sentence the narrator asserts a proposition about the fictive world of the text that is incorrect in that world. So to find out whether a sentence is a case of incorrect assertion, we need to know which proposition(s) about the fictive world the narrator asserts in it and what is true in the fictive world concerning the matter of those propositions.

The question of how we can find this out, however, seems to be a little more difficult. While part of the answer may be found in the definition, much of it depends on how exactly the phenomenon is implemented in the specific literary corpus, text or passage under study, or, more precisely, on the ambiguity or indeterminacy of the text with respect to the relevant questions. Thus, there may often be no general answer to this question – instead, the answer varies, for example, between genres or types of text,

---

9. As I will explain in section 5, we are, however, not aiming at a complete operationalization for computers in CAUTION, but only for humans. Due to its relatively high degree of interpretation dependence, only a partial operationalization for computers is realistic.

10. Another way to put this would be to say that the notion of interpretation dependence is meant to explain *de re* disagreement between scholars about fictional texts, not *de dicto* disagreement, i.e. disagreement about the object in question instead of about how we should talk about it.

or even between individual texts and passages. So the best we can do may be to find parameterized answers that together cover most or many cases, which may lead to the identification of new, application-related subtypes of the phenomenon.[11]

To illustrate this, let's look at unreliable narration/incorrect assertion. The first question would be how to find out which proposition the narrator is asserting in a sentence. This may seem straightforward – but in a surprisingly large number of cases, phenomena such as incomplete sentences, rhetorical questions, syntactic or lexical ambiguity, as well as irony or figurative speech can make it difficult to identify which statement a narrator makes at a particular point in the text. To formulate rules for identifying propositions we would probably need to refer to, for example, Gricean conversations maxims (Grice 1975)[12] – but determining the appropriate level of detail when specifying rules for different subcases is a very challenging task.

The second question that needs to be answered in order to detect unreliable narration/incorrect assertion seems even more challenging: How do we find out what is true in the fictive world of a text? While there are theories and methods for answering this question, such as the reality principle or the principle of minimal departure (Lewis 1978; Ryan 1980), these theories do not seem to be of much help for many of the questions about fictive worlds that are typically relevant in the context of unreliable narration. For example, the reality principle/principle of minimal departure suggests that a fictive world resembles our world as much as possible, meaning that unless the text suggests otherwise, we should 'import' the knowledge we have about our world into the fictive world. However, in many cases of potentially unreliable narration, we have to choose between alternative possible events that could just as well happen in our world.[13] On the other hand, if the choice is between a realistic and a supernatural possibility, the text often supports (roughly) equally well the different readings (an example would be Hoffmann's *Der Sandmann* ([1816] 1994)); and if we are dealing with a world that contains supernatural elements to begin with, we cannot be sure of the extent to which the fictive world differs from ours. In general, finding out what is true in the fictive world of a text will require a complex consideration and selection of various textual features and extratextual assumptions and information, and writing this down as rules is far from trivial as we would again need parameterized or even individualized sets of rules, which seems hardly feasible.

This is a point where we need to be particularly careful when operationalizing complex concepts: Because it is often so difficult to pin down the steps that are necessary to decide whether a concept applies, there sometimes seems to occur a shift between steps 1 and 2 of operationalizing a concept – especially in the context of computational literary studies. Based on the assumption that some of the relevant features may be indeterminable or immeasurable, the 'translation' of a definition into a step-by-step measurement guide actually results in a new, *pragmatic* definition the extension of which

---

11. This may require the operationalization of other concepts, such as genre categories.

12. The idea of using Gricean maxims to operationalize unreliable narration has been taken up by Heyd (2006, 2011) and by Kindt (2008, 53–67) (albeit with a special twist: The narrator violates the cooperation principle, while the author adheres to it). However, both approaches seem to start from a higher level than the one proposed here.

13. This is the case, for example, in Schnitzler's novella *Andreas Thameyers letzter Brief* (1961): Whether the narrator's wife has sexually betrayed her husband or been raped is not a question of proximity to our real world.

is (probably) not identical to the extension of the original one and can be regarded as an approximation to the original concept. This practice is even more common when the addressee of this operatIonalization step is a computer rather than a human being (Pichler and Reiter 2021, 5), as the integration of non-truth-preserving inferences and extratextual assumptions (see below) in explainable computational models seems to pose a particular challenge.This kind of reductionism is problematic for reasons I have laid out in section 3 above: It increases the gap between traditional and computational literary studies.

In contrast to the issues associated with the definition of a concept (see section 3 above), the challenges that arise in trying to identify the steps necessary to determine whether a concept actually applies seem to be a consequence of the high degree of interpretation dependence of the concept and related statements about literary texts. To substantiate this claim – and to generalize from it – I will now talk about interpretation dependence in more detail, elaborating on some of the concept's aspects that have been mentioned in subsection 2.2. As I have explicated there, interpretation dependence in the context of literary studies is understood here as a gradable feature of statements about literary texts (and of concepts used to make these statements).

Let us first consider statements about literary texts. A statement's *degree of interpretation dependence* is determined by two factors (which may in turn be present in different degrees): its reliance on (controversial) extratextual assumptions and its reliance on non-truth-preserving inferences. To say that a statement *relies* on something presupposes the assumption that statements about literary texts are typically arguable, i.e. that we can provide arguments to support them (Descher and Petraschka 2019). An argument typically consists of premises and conclusions, and we use some kind of inference rule to get from the premises to the conclusion. Typically, there are a number of arguments that directly support a claim – and often it makes sense to also provide arguments that support individual premises, and so on. Thus the first factor determining the degree of interpretation dependence of a statement – its reliance on (controversial) extratextual assumptions – is a feature concerning the premises of the arguments supporting a statement. If it relies solely on textual features, this aspect factors into its degree of interpretation dependence with zero. If extratextual assumptions are required to justify the statement, the degree of this aspect of interpretation dependence is determined by the number of extratextual assumptions required, the number of controversial extratextual assumptions and their degree of controversy, as well as their relevance in the argumentation. The second factor that determines the degree of a statement's interpretation dependence – its reliance on non-truth-preserving inference – is a feature that concerns not the premises but the inference rules/inferences used to get from the premises to the conclusion. If a statement can be deduced from the premises that support it, then this aspect factors into its interpretation dependence with zero.[14] If non-truth-preserving inferences (such as induction or inference to the best explanation) are required, the degree of this aspect of interpretation dependence is determined by the number of

---

14. As Descher (2019) has shown, deductive reasoning is in fact used in some arguments supporting interpretation hypotheses. If we take a closer look at his examples, it becomes clear that these deductive arguments, if used in the context of interpretations, contain at least one premise that is clearly in need of further argumentative support. We could either stop here and classify those premises as extratextual/controversial. Or we could work on filling in the argumentative gaps and would probably come across elements that will increase the degree of interpretation dependence.
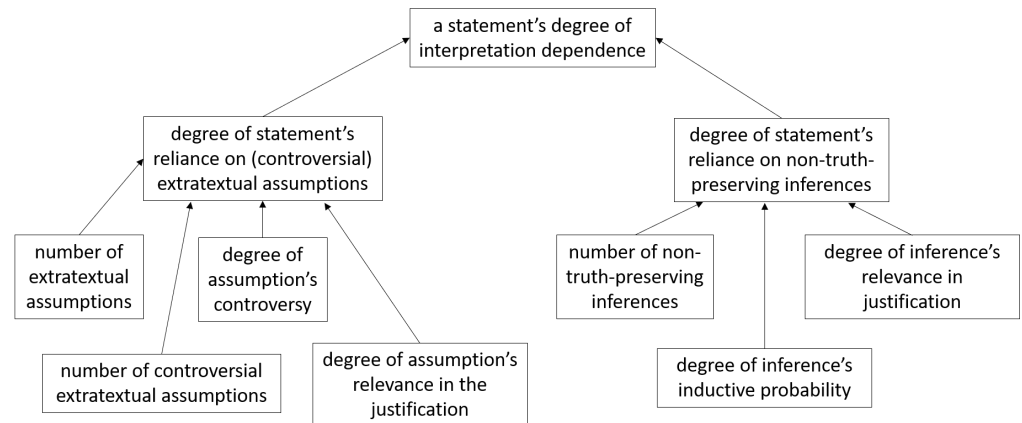
Figure 1: Factors influencing a statement's degree of interpretation dependence.

such inferences, their inductive probability (lower probability means higher degree of interpretation dependence) and their relevance in the argumentation (cf. Figure 1).

Now, none of the factors that make up the degree of interpretation dependence are usually easy to determine, and most of them probably can't be determined exactly. In particular, aspects such as relevance or degree of controversy will often be a matter of individual judgment. It is also difficult to stipulate a way how the factors should be weighed against each other. If a statement about a literary text is based solely on textual features and deductive reasoning, it is not dependent on interpretation. However, depending on how many and to what extent the different aspects contribute more than zero to the interpretation dependence, the degree will vary between slightly above zero and an open maximum. But even if an exact calculation of the degree of interpretation dependence may not be possible (or not yet possible, see section 6 below), knowing the factors that determine it will help us to estimate its degree for certain statements and concepts and thus to understand operationalization problems better.

In addition to the degree of interpretation dependence, we can distinguish between different *types of interpretation*[15] on which a statement about a literary text can be based. The first criterion constituting types of interpretation is the extent of the linguistic unit being interpreted – so basically the question is whether an interpretation applies only to a small part of a text (such as a sentence or even a word) or to the text as a whole. While this criterion, again, allows for a graded scale, we could call its poles 'micro-' and 'macro-level interpretation'. Although in practice it can be difficult to determine the extent of the linguistic unit to which an interpretation applies, there seem to be relatively clear cases. To take up unreliability-related examples from above, it is a micro-level interpretation to decide whether a single statement by the narrator is meant ironically – even if some cotext is usually needed to make this decision –, but it is a macro-level interpretation to reconstruct the plot-building events in the fictive world the text is about. Because macro-level interpretations are usually far more complex than micro-level ones, relying on more assumptions and inferences, statements based on them tend to show a higher degree of interpretation dependence.

15. I am understanding "interpretation" here as the result of attributing non-lexical aspects of meaning to (parts of) a literary text.

segment size / int. endeavor	micro-level ⟵⟶ macro-level
content-specifying	
content-transcending	

Figure 2: Types of interpretation.

A second criterion that constitutes types of interpretation is the larger-scale interpretive endeavor (and the associated types of meaning) of which the statement about a text is a part. Interpretations of literary texts are very often either aimed at reconstructing aspects of the fictive world or something beyond it, e.g. instances of figurative meaning located not on the intrafictional communication axis (between narrator and their addressee) but on the extratextual axis of communication (between author and reader, cf. Schmid 2008, 48). These types of interpretation have previously been labeled 'content-specifying' and 'content-transcending interpretation' (Folde 2015). Both types can occur at both the micro and macro levels (cf. Figure 2). The previously given examples illustrate content-specifying interpretation on the micro and macro level respectively. An example of micro-level content-transcending interpretation would be the attribution of some kind of figurative meaning to a text/story element,[16] while a macro-level interpretation of this type would involve deriving a generalized message from a text as a whole.[17] Content-transcending interpretation tends to be more heavily interpretation-dependent than content-specifying interpretation, if only for the reason that the latter type usually relies heavily on text-based arguments (Tepe et al. 2009), while the former type relies on theories of interpretation and related contextual information/assumptions,[18] which are also controversial in many cases.[19]

Having discussed degree and type, the final aspect of interpretation dependence that I would like to comment on is the *reason why a statement about a literary text can be interpretation-dependent*. One possibility is that this reason is linked to a literary studies concept that is used to make a statement about a text. Variants of this reason are that the definition of the concept refers to elements that are necessarily related to interpretation dependence (e.g. when it refers to controversial contextual assumptions)[20] or that the definition refers to elements whose identification often or usually depends on non-truth-preserving inference or (controversial) textual assumptions. An example of the latter is the reference to the fictive world in the definition of unreliable narration/incorrect assertion: As a content-based (rather than form-based) concept, reconstructing the fictive world of a text often (but not necessarily) involves interpretation. Many concepts in literary studies have such a minimum degree of, or tendency towards, interpretation dependence. Narratological time categories, for example, usually require a reconstruction of the fictive events, their chronology and duration. Exceptions are some purely formal

---

16. For example when Klara in *Der Sandmann* is read as the embodiment of reason.

17. An example would be to assume that *Andreas Thameyer's letzter Brief* communicates a critical message about the fragility of the male ego.

18. For example, if we want to know the message of *Der Sandmann*, we need to decide whether we should try to reconstruct the author's intentions to answer this question, whether we can consult sketches and earlier versions of the text, etc.

19. It is possible, however, that arguments in favor of content-specifying interpretations contain content-transcending interpretation hypotheses.

20. An example would be Booth's definition of unreliable narration, which refers to the implied author, whose 'existence'/fruitfulness is a highly controversial assumption (Booth 1959, 58–59).
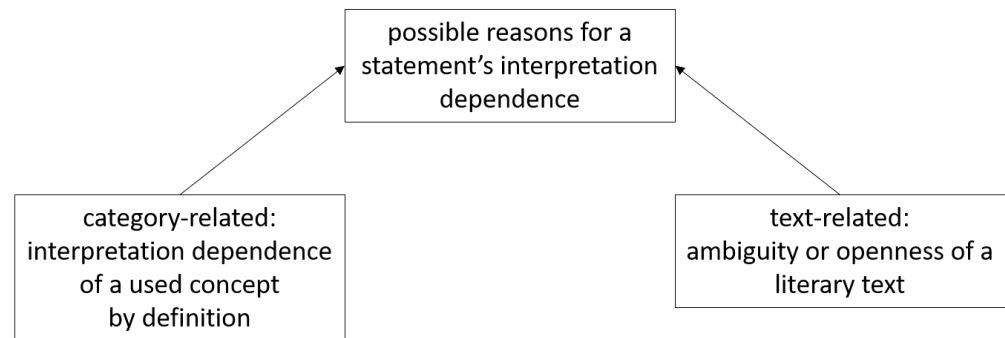
Figure 3: Reasons for interpretation dependence.

or stylistic concepts, some of which may even preclude interpretation dependence, e.g. if we analyze a text according to how colloquial its language is.

The second possibility is that the reason for interpretation dependence does not lie in the definition of the concept itself but in the makeup of the literary text, especially its potential ambiguity or openness (with regard to the phenomena in question). Some of the seemingly least interpretive concepts require interpretation when they are used to make statements about a text that is open or ambiguous in the relevant respects (cf. Figure 3).[21] Again, this aspect may vary according to individual texts as well as individual text types. The example of unreliable narration/incorrect assertion is particularly interesting in this case: Typically, texts in which the narrator is potentially unreliable are constructed in such a way that makes statements about the fictive world (and thus statements about unreliable narration) highly dependent on interpretation. The particular type of text that falls into the category of literary fantasy is even defined by difficulties in identifying the relevant aspects of its world (oscillating between natural and supernatural readings, cf. Todorov 1972). On the other hand, however, in texts with a certain (application-related?) type of unreliable narration – namely resolved unreliability –, it seems not (or very little) interpretation-dependent to make the relevant diagnoses about the fictive world.[22]

So, overall, unreliable narration/incorrect assertion seems to be a relatively high-degree interpretation-dependent concept (and statements about this type of unreliability of a narrator in a literary text are also relatively high-degree interpretation-dependent). This interpretation dependence seems to be the main reason why it is so difficult to identify the steps necessary to detect unreliability in a text. To formulate some general tendencies: Such problems seem to affect to a greater extent cases where there is a high degree of interpretation dependence, where the interpretation dependence is rooted in the definition of a concept and/or a (corpus of) highly ambiguous texts are analyzed, and where macro-level (especially content-transcending) interpretation is necessary.

How did we deal with the challenges posed by the relatively high degree of interpretation dependence of unreliable narration in CAUTION in the context of identifying the steps

---

21. For example, narratological time categories are usually considered non- or low-degree-interpretive. However, when combined with a potentially unreliable narration, the classification of text passages e.g. according to their narrative speed may depend on whether the narrated events are read as a dream or not, as in Eichendorff's *Auch ich war in Arkadien* ([1866] 2012).
22. An example of resolved unreliability is Frisch's *Stiller* ([1954] 2001), in which another narrator confirms in an epilog that the main narrator is in fact Anatol Ludwig Stiller, which the latter consistently denies.

necessary to detect it? In general, when operationalizing a concept, the three different tasks (definition, identification of steps, decision making) are performed iteratively and inform each other (Gius and Jacke 2017 and Reiter 2020 for annotation-based operationalization). In CAUTION, this intermingling is particularly true for the second and the third tasks – and the idea of optimizing the step-by-step description of how to identify unreliable narration in a practice-based approach, i.e. by trying to make a (well-considered) decision about whether the concept applies, actually inspired a large part of the conceptual setup of the project, as I will explain in the following section. While our project to operationalize unreliable narration/incorrect assertion was not concluded yet at the time of writing this paper, I will give a few examples of how our attempt to decide whether the concept applies helped to refine the steps necessary for identifying it.

5. Deciding Whether a Concept Applies

CAUTION is designed as a three-tracked project, experimenting with different general approaches to operationalizing unreliable narration/incorrect assertion (or a reasonable approximation). In the following, I will introduce the core track, the deep track and the approximation track, explain how they interact and how they may help to gain further insights concerning the steps necessary for identifying our variant of unreliable narration.

The *core track* deals with unreliable narration as defined above in a 'conventional' annotation approach. While we do use annotation guidelines for this, the guidelines not yet attempt to detail the (parameterized) steps for identifying which proposition a narrator is asserting and how to reconstruct the fictive world. Instead, they mainly specify what an asserted proposition is (e.g. distinguishing it from evaluative utterances), which passages should generally be excluded from annotation (e.g. embedded speech), and how to select the annotation span (e.g. mainly subsentences with predicate, sometimes noun phrases).[23] Human annotators have been asked to annotate texts from a small corpus of German fictional narratives from the 19th century to the present. The corpus consists of four narratives that are usually considered unreliable (with different estimated degrees of ambiguity), four narratives that feature related phenomena (like improbable events or a highly personalized narrator) and one narrative without such features. The task is to annotate (sub)sentences in which the narrator makes an incorrect claim about the fictive world ("incorrect statement") or might make such an incorrect claim ("undecided"). So the idea here is to operationalize unreliable narration as a category for text analysis. In addition to this, we experiment with using the concept to describe texts or narrators rather than individual (sub)sentences, as this seems to be the way concepts are often used in literary studies practice. Here, the annotators are asked to provide both a graded and an (almost) binary label for each text (in percentage and as "yes"/"no"/"undecided"-decision respectively). The annotation guidelines deliberately do not specify how such a label should be arrived at (apart from emphasizing that it should not be calculated simply on the basis of the number of annotations in the text). Also, the different types of annotations or categorizations are not the subject of discussion among annotators in this core track. The idea behind this is to collect

23. The current version of the guidelines can be found here: Blessing et al. 2024b.

some descriptive data on how such decisions turn out in practice – in the hope that data analysis will provide insight into possible regularities and relations. To increase the chances of success, we are also generating other types of data, as I will explain below.

In an attempt to optimize the annotation guidelines, i.e. the identification of steps for determining the target phenomenon as a part of operationalizing the concept for human addressees, we designed the *deep track*. It is aimed at analyzing the previous annotation decisions by, metaphorically speaking, taking a step back from the close reading and annotation of the texts and trying to develop and argue for a macro-level content-specifying interpretation for each text. In other words, the idea is to arrive at a coherent and well-founded reconstruction of the most relevant aspects of the fictive world for each text – which is one of two relevant steps for making the decision of whether unreliable narration in the form of an incorrect assertion occurs, as I have explained in section 4. In order to structure and systematize this complex task, and to make the results comparable, it is divided into the following subtasks: (1) identifying the relevant open questions about the fictive world, (2) identifying the possible competing answers to these questions, (3) selecting a preferred answer per question and (4) providing arguments for these decisions that are organized in the form of argument trees (Descher et al. 2023; Winko et al. 2024), for which visualization software such as MindMup[24] or Argdown[25] can be used. This task differs conceptually from annotation in that the data that is being generated is initially not directly linked to offsets of the text. Instead, what is being constructed is a proposition-based model of the text (or an important aspect of the text, namely the world it describes), which is systematically enriched with arguments that support the decisions leading to that model. Annotations (or offsets of the relevant text) may come back into play in the form of premises of arguments in the model, since (parts of) the text itself will be among the most important premises supporting content-specifying interpretation hypotheses of the text.

The connection between this experimental deep track and the overall goal of operationalizing unreliable narration is that it is intended to provide insight into the most obscure step in identifying unreliable narration – namely the reconstruction of the fictive world. The aim here is to get the former annotators to document and reflect on the decisions that influence their annotations, which they may have previously made mainly intuitively. In addition, an evaluative facet comes into play: Not only does the required way of documenting and providing arguments promote coherence and consistency more than annotation – the specific preparation of interpretation hypotheses and arguments also allows for a systematic and informed discussion among the former annotators, which may lead to a reconsideration of decisions in some cases. Ideally, this step will lead to both better informed annotation decisions and better step-by-step guidelines for identifying unreliable narration.

At the time of writing this paper, the former annotators have developed and discussed argument trees for two of the project's corpus texts (Bendixen's *Meine falschen Eltern* (cf. Bendixen 2015) and Schnitzler's *Andreas Thameyers letzter Brief*, which were estimated to exhibit a low resp. medium degree of ambiguity with regard to the relevant questions

---

24. See https://www.mindmup.com/.
25. See https://argdown.org/.

about the fictive world) and revised their formerly intuitive annotation decisions concerning the occurrence of incorrect assertions on this well-reflected basis. The argument trees have not yet been systematically evaluated to extract more detailed instructions on how to reconstruct the relevant parts of a fictive world. However, here are some potentially interesing observations. First, for all trees (across texts and annotators), partial summaries or paraphrases of narrated content play important roles as premises,[26] and these summaries/paraphrases are in turn supported argumentatively by providing quotes from the texts.[27] The annotators tend to largely agree on which paraphrases and text passages are relevant in this context, especially on the most basic level(s) of the argumentation. Second, another common type of premise used to support arguments about the fictive world are psychological hypotheses. These are typically used to argue for hypotheses about characters' motives for acting in a certain way. It is noticable here that annotators not only tend to cite different psychological literature – sometimes they also refer to different psychological mechanisms when explaining characters' behavior.[28] Third, apart from psychological assumptions that serve as premises, annotators tend to base their arguments on different kinds of extratextual information,[29] the relevance of which is sometimes argued for by referring to theories of interpretation.[30] However, there were no cases in the argument trees in which theories of interpretation have in turn been supported argumentatively.

While a systematic evaluation of how exactly such results may help to refine the description of the steps necessary to identify unreliable narration in the form of incorrect assertion has to occur elsewhere, I would briefly like to mention two possibilities. First, if it turns out that one premise type occurs frequently and with high agreement between annotators, it could be advised to draw upon this type in the guidelines. Second, better knowledge about the possible contexts that are drawn upon when arguing for a reconstruction of a fictive world may help to stipulate specific contexts for a precise and narrower operationalization[31] or to parameterize operationalizations.

The two tracks I presented so far are mainly aimed at operationalizing the concept of unreliable narration for human addressees. In contrast, the main aim of the *approximation track* is a computational operationalization of 'something like' unreliable narration – resulting in the automated detection of text features that are very often associated with unreliable narration. This track is based on the fact that literary studies research on unreliable narration often compiles lists of indicators that may hint that a text being unreliably narrated (Allrath 1998; Nünning 1999). These lists are usually very heterogeneous, ranging from linguistic text properties (e.g. exclamations) to narrator characteristics (e.g. emotional agitation). For our project, we have selected six potentially indicative narrator characteristics, some of which are assumed to correlate with specific linguistic

26. E.g. "Thameyer claims his only reason for wanting to commit suicide is his love for his wife" for Schnitzler's *Andreas Thameyers letzter Brief*.
27. E.g. "Es ist ja nur aus Liebe zu dir, daß ich sterbe" (It is only out of love for you that I am dying, J.J.), Schnitzler 1961.
28. For example, annotators refer to the phenomenon of cognitive dissonance vs. psychosis to explain Thameyer's behavior.
29. This includes information about psychological assumptions known to be held by the author of the text, about earlier versions of the text, about actual events that may have inspired the fictive events narrated in the text etc.
30. For example, reference to the theory of intentionalism has been used to justify drawing upon knowledge about the author's beliefs about psychology.
31. This has, for example, been done by Pichler and Reiter (2021) when operationalizing the concept of mysteriousness according to the hermeneutic approach in the variant advocated by Altenhofer.

text properties. Some of these indicators can already be detected relatively reliably with computational models (such as indicators of emotion (Klinger et al. 2020) – even though the underlying definition may differ in detail from the relevant unreliability indicators), others seem to be comparably easy to model, e.g. even in a rule-based approach (such as addressee orientation/awareness of a communicative situation). The idea is to apply such automated models to the corpus and to let human annotators detect them manually. This will not only allow a comparison between an automated and a manual detection of these indicators (as well as a possible refinement of the models based on the annotations) – it will also make it possible to analyze the actual relation of the indicators to unreliable narration/incorrect assertion in two ways (Jacke 2023b): by explaining whether and how the relevant indicators are logically related to unreliable narration, and by analyzing whether and to what extent the indicators actually co-occur with the detection of unreliability by the annotators. So far, we have concluded the manual annotation of the indicators and we have conducted first analyses of the relationship between manually annotated indicators and manually annotated cases of unreliable narration, which suggest that the selected linguistic indicators occur slightly more often in sentences that are categorized as incorrect assertion (Blessing et al. 2024a; Blessing et al. 2024b) – but further analyses are needed to substantiate and differentiate this claim. If successful, it may be possible to include references to certain indicators in the guidelines on how to determine the relevant aspects (uttered propositions and corresponding features of the fictive world).

The intended outcome of the approximation track is to develop a computational model that can automatically identify textual features that are likely to co-occur with unreliable narration. Another way of putting this would be to say that the approximation track is based on an alternative/pragmatic definition of a different concept of unreliability the extension of which is not identical to the concept of unreliable narration that is of direct interest to literary studies. The aim would be to use the data generated in the other tracks of CAUTION to analyze and explain as precisely as possible how the automated model relates to the original concept – this is a prerequisite for the model to be useful in literary studies contexts. In contrast to the other two tracks, the aim here is a (partial or approximate) operationalization not for humans but for computers, and it will be accompanied by instructions on how to work on with it in human operationalization, analysis and interpretation efforts.[32]

In the following and final section of this paper, I will summarize the main points of the paper and attempt a generalization from the solutions adopted in CAUTION for operationalizing highly interpretation-dependent concepts and briefly sketch out ways of evaluating the different elements suggested in this paper.

6. Conclusion

This paper's aims were to show how the operationalization of literary studies concepts may be complicated by a high interpretation dependence of these concepts and to suggest ways of addressing these problems to bridge the gap between traditional and

---

32. It is interesting (and potentially challenging), however, that, as first analyses of the annotation data have shown, the inter-annotator agreement seems to be higher for the detection of unreliable narration that for the so-called indicators. Possible explanations are discussed by Blessing et al. (2024a).

computational literary studies. The theoretical ideas have been illustrated by drawing upon an example case: the ongoing attempt to operationalize the concept of unreliable narration. After having laid out the basics (definitions of "operationalization" and "interpretation dependence" as well as the interplay between theory and practice in connection with this paper's aims), the paper has followed the steps of operationalizing literary studies concepts (i.e. defining the concept, identifying the steps, making the decision) with a focus on the implication of interpretation dependence in this process. The first step (that of developing a definition) is particularly difficult with literary studies concepts, as they typically aim to capture very complex and often heterogeneous phenomena and are neither introduced nor used in a standardized way. These problems are not directly related to interpretation (i.e. the act of making sense of a linguistic entity) but instead to conceptualizing. One suggestion for dealing with complex concepts is to select and weigh quality criteria for defining literary studies concepts and to make this reasoning explicit. In the case of very heterogeneous concepts, it is often useful to work with selected sub-concepts.

The second step of operationalizing literary studies concepts is to identify the steps necessary to determine whether the concept applies to a textual entity, which can be divided into two sub-steps: identifying what needs to be known in order to make this decision, and explaining how these things can be found out. While the first sub-step can often be carried out on the basis of the concept's definition, the second is often very difficult and needs to be parameterized according to text types or even individual texts. Its difficulty seems to increase with the degree of interpretation dependence of the intended statements to be made about texts. The degree of interpretation dependence is determined by the extent to which a statement relies on (controversial) extratextual assumptions (taking into account the number and relevance of these assumptions plus the degree of controversy) or on non-truth-preserving inferences (taking into account the number, relevance and degree of inductive probability of these inferences). This degree increases when the concepts used to make a statement about a text are interpretation-dependent by definition, when the texts are ambiguous, when the statement concerns the text as a whole and aspects of meaning that go beyond a reconstruction of the fictive world.

While interesting ways of dealing with moderately interpretation-dependent concepts have been suggested (Pichler and Reiter 2022), a first recommendation of dealing with highly interpretation-dependent concepts (and their application to ambiguous texts) in the context of operationalization is to not try to strictly separate the second and third steps of operationalization (i.e. identifying the steps and performing the decision) but to use a collaborative and discursive setting of applying the concept/performing the decision as a means of gaining further insight into the often obscure steps. A second recommendation is not to attempt a full operationalization for computers but to go with a deliberate approximation/partial operationalization and to explain its relation to the original concept as precisely as possible.

Both of these recommendations are based on the acceptance that a full (and interpretable) computational operationalization of highly interpretation-dependent concepts is currently not possible. However, a cautious and modest approach to this issue from several sides, as suggested here, might make a full computational operationalization of

highly interpretation-dependent concepts possible in the future: As many concepts – as well as the theories, methods and practices of interpretation – in literary studies are complex and opaque, most steps towards operationalizing highly interpretation-dependent concepts require more than just theoretical efforts. As suggested here, carefully designed practice-based collaborative studies, in which the assumptions and processes involved in applying a concept are thoroughly documented and deliberated upon, can pave the way to a better understanding of the relevant practices – which, in turn, is a prerequisite for the development of understandable computational models that serve the interests of literary studies.

To conclude this paper, I will briefly touch upon the question of how the most important theses and suggestions I developed in this paper can be evaluated. First, my suggested definitions of "operationalization" and "interpretation dependence" can be challenged by convincingly arguing that they don't meet the quality criteria that I aimed to meet (i.e. sufficient degree of similarity with previous uses of the relevant terms, increased exactness, fruitfulness for traditional literary studies purposes), that the quality criteria should be weighed differently or that entirely different quality criteria should be aimed for. If my definitions are refuted, my argument is likely significantly weakened. Second, the usefulness of my analysis of the relation between interpretation dependence and operationalization problems can be questioned by showing that interpretation dependence can't in fact explain the operationalization problems it claims to explain, or by showing that there are significant operationalization problems that can't be explained via interpretation dependence.[33] Third, my suggestions for dealing with the operationalization problems can be questioned by arguing either that the suggestions are not feasible or that they don't mitigate the problems they claim to mitigate. Forth and finally, how about the operationalization of unreliable narration I'm sketching out in this paper? As I spelled out in subsection 2.3, this specific operationalization attempt mainly serves to illustrate my theoretical thoughts but ultimately does not play a crucial part in justifying them. Also, the operationalization is difficult to assess because it is not finished yet. One thing I would like to point out, however, is that typical evaluation methods for operationalizations for humans or computers (like inter-annotator agreement resp. precision, recall, F1 score) in their current form have limited significance when it comes to highly interpretation-dependent concepts. Instead, a well-founded and discursively examined argumentation justifying annotation decisions may be a (complexity- and ambiguity-, hence literary studies-friendly, yet very time-consuming) way to go.[34] There is still some work left to do here.

7. Data Availability

As this is a theoretical contribution to the field of computational literary studies, this article does not rely on specific data. However, annotation data from the example project CAUTION used to illustrate some of its points is provided in connection with a previous project-related publication. The data can be found at: https://doi.org/10.5281/zenodo.10254506.

---

33. In the latter case, however, it may be possible to merely supplement my theory with additional explanations for other operationalization problems.
34. For the idea to establish a new kind of standard, namely platinum-standard annotations, based on this approach, cf. Jacke (2025).

8. Acknowledgements

9. Author Contributions

Janina Jacke: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing

References

Allrath, Gaby (1998). "'But why *will* you say that I am mad?' Textuelle Signale für die Ermittlung von *unreliable narration*". In: *Unreliable Narration. Studien zur Theorie und Praxis unglaubwürdigen Erzählens in der englischsprachigen Erzählliteratur*. Ed. by Ansgar Nünning, Carola Surkamp, and Bruno Zerweck. VWT, 59–80.

Bendixen, Katharina (2015). "Meine falschen Eltern". In: *Gern, wenn du willst*. Poetenladen, 27–34.

Blessing, André, Janina Jacke, and Jonas Kuhn (2024a). "Agreement und Kookkurrenz bei unzuverlässigem Erzählen. Ziele, Herausforderungen und erste Ergebnisse aus dem Projekt CAUTION". In: *DHd 2024. Quo Vadis DH. Konferenzabstracts*. Ed. by Joëlle Weis, Thomas Haider, and Estelle Bunout, 107–111.

— (2024b). *CAUTION Annotations*. [Dataset]. 10.5281/zenodo.10254506.

Booth, Wayne (1959). *The Rhetoric of Fiction*. Chicago University Press.

Brennan, Andrew (2022). "Necessary and Sufficient Conditions". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Metaphysics Research Lab, Stanford University.

Bridgman, Percy (1954). *The Logic of Modern Physics*. Macmillan.

Bühler, Axel (2003). "Die Vielfalt des Interpretierens". In: *Hermeneutik. Basistexte zur Einführung in die wissenschaftstheoretischen Grundlagen von Verstehen und Interpretation*. Ed. by Axel Bühler. Synchron, 99–119.

Carnap, Rudolf (1959). *Induktive Logik und Wahrscheinlichkeit*. Spinger.

— (1965). *Meaning and Necessity. A Study in Semantics and Modal Logic*. Chicago University Press.

Cook, Roy T. (2009). "Intensional Definition". In: *A Dictionary of Philosophical Logic*. Edingurgh University Press, 155.

Descher, Stefan (2019). "Deduktive Schlüsse in der literaturwissenschaftlichen Praxis". In: *Journal of Literary Theory* 13 (2), 145–160. 10.1515/jlt-2019-0005.

Descher, Stefan, Jan Borkowski, Felicitas Ferder, and Philipp David Heine (2015). "Probleme der Interpretation von Literatur – Ein Überblick". In: *Literatur interpretieren. Interdisziplinäre Beiträge zur Theorie und Praxis*. Ed. by Jan Borkowski, Stefan Descher, Felicitas Ferder, and Philipp David Heine. Mentis, 11–70.

Descher, Stefan, Merten Kröncke, and Simone Winko (2023). "Wie plausibilisieren Literaturwissenschaftler*innen ihre Interpretationen? Das DFG-Projekt 'Das Herstellen von Plausibilität in Interpretationstexten. Untersuchungen zur Argumentationspraxis in der Literaturwissenschaft' (ArguLit)". In: *Textpraxis Sonderausgabe* 7 (2), 1–7. 10.17879/19958496834.

Descher, Stefan and Thomas Petraschka (2019). *Argumentieren in der Literaturwissenschaft. Eine Einführung*. Reclam.

Eichendorff, Josef von [1866] (2012). *Auch ich war in Arkadien*. TextGrid Repository. https://textgridrep.org/browse/msfg.0 (visited on 02/04/2025).

Folde, Christian (2015). "Grounding Interpretation". In: *The British Journal of Aesthetics* 55 (3), 361–374. 10.1093/aesthj/ayv020.

Fricke, Harald (1970). *Die Sprache der Literaturwissenschaft. Textanalytische und philosophische Untersuchungen*. Beck.

Frisch, Max [1954] (2001). *Stiller*. Suhrkamp.

Genette, Gérard (2010). *Die Erzählung*. 3rd ed. Fink.

Gerstorfer, Dominik (2020). "Entdecken und Rechtfertigen in den Digital Humanities". In: *Reflektierte algorithmische Textanalyse*. Ed. by Nils Reiter, Axel Pichler, and Jonas Kuhn. De Gruyter, 107–124. 10.1515/9783110693973-005.

Gerstorfer, Dominik and Evelyn Gius (2025). "Operationalizing operationalizing". In: *DHd 2025. Under Construction. Konferenzabstracts*. Ed. by Nils Reiter, Thomas Haider, Daniel Kababgi, and Hendrik Buschmeier, 345–348. 10.5281/zenodo.14943080.

Gius, Evelyn (2016). "Narration and Escalation. An Empirical Study of Conflict Narratives". In: *DIEGESIS* 5 (1). https://www.diegesis.uni-wuppertal.de/index.php/diegesis/article/view/222 (visited on 02/04/2025).

— (2019). "Computationelle Textanalysen als fünfdimensionales Problem: Ein Modell zur Beschreibung von Komplexität". In: *LitLab Pamphlets* 8. Ed. by Thomas Weitin. https://www.digitalhumanitiescooperation.de/wp-content/uploads/2019/12/pamphlet_gius_2.0.pdf (visited on 02/04/2025).

Gius, Evelyn and Janina Jacke (2017). "The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis". In: *International Journal of Humanities and Arts Computing* 11 (2), 233–254. 10.3366/ijhac.2017.0194.

— (2022). "Are Computational Literary Studies Structuralist?" In: *Journal of Cultural Analytics* 7 (4). 10.22148/001c.46662.

Grice, Paul Herbert (1975). "Logic and conversation". In: *Syntax and semantics*. Ed. by Peter Cole and Jerry Morgan. Vol. 3: Speech acts. Academic Press, 41–58.

Heyd, Theresa (2006). "Understanding and handling unreliable narratives. A pragmatic model and method". In: *Semiotica* 162, 217–243. 10.1515/SEM.2006.078.

— (2011). "Unreliability. The Pragmatic Perspective Revisited". In: *Journal of Literary Theory* 11 (1), 3–17. 10.1515/jlt.2011.003.

Hoffmann, E. T. A. [1816] (1994). "Der Sandmann". In: *Gesammelte Werke in Einzelausgaben*. Vol. 3. Aufbau-Verlag.

Jacke, Janina (2019). *Systematik unzuverlässigen Erzählens. Analytische Aufarbeitung und Explikation einer problematischen Kategorie*. Narratologia 66. De Gruyter. 10.1515/9783110659689.

— (2023a). "Die (computationelle?) Operationalisierung unzuverlässigen Erzählens. Ein Beitrag zur Theorie und Methodik literaturwissenschaftlichen Interpretierens". In: *Textpraxis* 21 (2). 10.17879/19958499445.

Jacke, Janina (2023b). "Vom sprachlichen Indikator zum komplexen Phänomen? Operationalisierungsprobleme in der computationellen Literaturwissenschaft am Beispiel des unzuverlässigen Erzählens". In: *DHd 2023. Open Humanities, Open Culture. Konferenzabstracts*. Ed. by Anna Busch and Peer Trilcke, 317–321.

— (2025). "Platinstandard-Annotation in der digitalen Literaturwissenschaft: Definition, Funktionen und diskursive Argumentvisualisierung als Best-Practice-Beispiel". In: *DHd 2025. Under Construction. Konferenzabstracts*. Ed. by Nils Reiter, Thomas Haider, Daniel Kababgi, and Hendrik Buschmeier, 279–283. `10.5281/zenodo.14943180`.

Kindt, Tom (2008). *Unzuverlässiges Erzählen und literarische Moderne. Eine Untersuchung der Romane von Ernst Weiß*. Niemeyer.

Kindt, Tom and Hans-Harald Müller (2003). "Wieviel Interpretation enthalten Beschreibungen? Überlegungen zu einer umstrittenen Unterscheidung am Beispiel der Narratologie". In: *Regeln der Bedeutung. Zur Theorie der Bedeutung literarischer Texte*. Ed. by Fotis Jannidis, Gerhard Lauer, Matías Martínez, and Simone Winko. De Gruyter, 286–304. `10.1515/9783110907018.286`.

Klinger, Roman, Evgeny Kim, and Sebastian Padó (2020). "Emotion Analysis for Literary Studies: Corpus Creation and Computational Modelling". In: *Reflektierte algorithmische Textanalyse*. Ed. by Nils Reiter, Axel Pichler, and Jonas Kuhn. De Gruyter, 237–268. `10.1515/9783110693973-011`.

Krautter, Benjamin (2022). "Die Operationalisierung als interdisziplinäre Schnittstelle der Digital Humanities". In: *Scientia Poetica* 26 (1), 215–244. `10.1515/scipo-2022-009`.

Kroon, Fred and Alberto Voltolini (2023). "Fictional Entities". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2023. Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/fall2023/entries/fictional-entities/` (visited on 02/04/2025).

Lewis, David (1978). "Truth in Fiction". In: *American Philosophical Quarterly* 15 (1), 37–46. `https://www.jstor.org/stable/20009693`.

Margolis, Eric and Stephen Laurence (2023). "Concepts". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2023. Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/fall2023/entries/concepts/` (visited on 02/04/2025).

Moretti, Franco (2013). ""Operationalizing": or, the function of measurement in modern literary theory". In: *Literary Lab Pamphlets* 6. `https://litlab.stanford.edu/LiteraryLabPamphlet6.pdf` (visited on 01/30/2025).

Nünning, Ansgar (1999). "Unreliable, Compared to What? Towards a Cognitive Theory of "Unreliable Narration". Prolegomena and Hypotheses". In: *Grenzüberschreitungen. Narratologie im Kontext / Transcending Boundaries. Narratology in Context*. Ed. by Walter Grünzweig and Andreas Solbach. Narr, 53–73.

Pichler, Axel and Nils Reiter (2021). "Zur Operationalisierung literaturwissenschaftlicher Begriffe in der algorithmischen Textanalyse. Eine Annäherung über Norbert Altenhofers hermeneutische Modellinterpretation von Kleists Das Erdbeben in Chili". In: *JLT Articles* 15 (1), 1–29. `10.1515/jlt-2021-2008`.

— (2022). "From Concepts to Texts and Back: Operationalization as a Core Activity of Digital Humanities". In: *Journal of Cultural Analytics* 7 (4). `10.22148/001c.57195`.

Reichert, John F. (1969). "Description and Interpretation in Literary Criticism". In: *The Journal of Aesthetics and Art Criticism* 27 (3), 281–292. 10.2307/428674.

Reiter, Nils (2020). "Anleitung zur Erstellung von Annotationsrichtlinien". In: *Reflektierte algorithmische Textanalyse*. Ed. by Nils Reiter, Axel Pichler, and Jonas Kuhn. De Gruyter, 193–202. 10.1515/9783110693973-009.

Ryan, Marie-Laure (1980). "Fiction, non-factuals, and the principle of minimal departure". In: *Poetics* 9 (4), 403–422. 10.1016/0304-422X(80)90030-3.

Schmid, Wolf (2008). *Elemente der Narratologie*. De Gruyter. 10.1515/9783110978520.

Schnitzler, Arthur (1961). "Andreas Thameyers letzter Brief". In: *Gesammelte Werke. Die erzählenden Schriften*. Vol. 1, 514–520.

Spree, Axel (2007). "Interpretation". In: *Reallexikon der deutschen Literaturwissenschaft*. Ed. by Harald Fricke. Vol. 2. De Gruyter, 168–172.

Tepe, Peter, Jürgen Rauter, and Tanja Semlow (2009). *Interpretationskonflikte am Beipsiel von E. T. A. Hoffmanns "Der Sandmann": kognitive Hermeneutik in der praktischen Anwendung*. Studienbuch Literaturwissenschaft 1. Königshausen & Neumann.

Todorov, Cvetan (1972). *Einführung in die fantastische Literatur*. Hanser.

Trilcke, Peer and Frank Fischer (2016). "Fernlesen mit Foucault? Überlegungen zur Praxis des *distant reading* und zur Operationalisierung von Foucaults Diskursanalyse". In: *Le foucaldien* 2 (1), 1–18. 10.16995/lefou.15.

Weimar, Klaus (2007). "Literaturwissenschaft". In: *Reallexikon der deutschen Literaturwissenschaft*. Ed. by Harald Fricke. Vol. 2. De Gruyter, 485–489.

Weitz, Morris (1964). *Hamlet and the Philosophy of Literary Criticism*. Meridian.

Winko, Simone, Stefan Descher, Urania Milevski, Merten Kröncke, Fabian Finkendey, Loreen Dalski, and Julia Wagner (2024). *Praktiken des Plausibilisierens: Untersuchungen zum Argumentieren in literaturwissenschaftlichen Interpretationstexten*. Göttingen University Press. 10.17875/gup2024-2639.