Evelyn Gius

Computational text analysis as a five-dimensional problem:

A model for the description of complexity

Pamphlet #8

August 2020

Ed. Thomas Weitin

DIGITAL
HUMANITIES
COOPERATION

Evelyn Gius

Computational text analysis as a five-dimensional problem:
A model for the description of complexity

Abstract

Computational approaches to literary studies, based on a number of both new and established procedures that enable computational text analysis in the interest of literary research, are now an integral part of Digital Humanities. Accordingly, there is a great need for description of and reflection upon these approaches, both within the Digital Humanities and with regard to the relationship between computational literary studies and non-computational literary studies. In an effort to facilitate such reflection, this paper presents a model that captures the complexity of computational text analysis, in relation to the phenomena and texts under consideration, as well as to the findings of the analysis. Specifically, five dimensions are proposed by which any computational text analysis in literary studies—and beyond—can be described: (1) the composition of the analyzed phenomena, (2) the contextualization of the phenomena, (3) the heterogeneity of the considered texts, (4) the mode of analysis, and (5) the cognitive contribution of computational analysis.

Bibliographic information of the Deutsche Nationalbibliothek: The Deutsche Nationalbibliothek lists this publication in the German National Bibliography. It is recorded in the Zeitschriftendatenbank (ZDB) and in the international ISSN portal. Detailed bibliographical data can be accessed on the Internet at http://dnb.d-nb.de.

ISSN: 2628-4537
Ed.: Thomas Weitin

Evelyn Gius

Computational text analysis as a five-dimensional problem:

A model for the description of complexity

1 Introduction

Digital Humanities projects are determined by a number of aspects that make them complex undertakings. These range from the negotiation of assumptions and methods and their fit to the subject matter; to the design of concrete interdisciplinary collaborations, which can be a great challenge for participants both personally and professionally, and in terms of their career strategies; to the presentation of results for one or more research communities. In addition to issues surrounding project planning and management, or strategic and scientific policy and communications, projects also deal with questions that affect their actual research processes. These latter questions are currently being discussed more and more in terms of their relevance and orientation. A harsh criticism by Nan Z. Da (2019a) of the procedures in Computational Literary Studies, a sub-field of literary Digital Humanities, initiated a debate called the "Digital Humanities War."[1] Among other things, this debate is presented as a confrontation between alleged structuralists and post-structuralists, wherein some self-defined non-structuralist literary scholars suggest an opposition and a gulf between structuralism and their own approaches (see, e.g., Dobson 2019 and Bode 2020). The debate is aggravated by the major role that funding policy plays in it, at least in its background.[2]

While this debate is important overall, the confrontational manner in which it is sometimes conducted detracts from the opportunity it would otherwise present to fully address the well-founded arguments of the respective opposing sides. From the point of view of Digital Humanities *and* Literary Studies, for example, it would be illuminating to describe procedures or even methods in more detail, and to reflect on their meanings before scrutinizing other aspects of the of the alleged "war." I thereby wish to propose a model for the consideration of approaches in computational text analysis, which, for the time being, allows a detailed look at a project independent of the above-described debates. This model serves as a rough estimate of the complexity of research approaches that apply computational analyses.

[1] Cf. the article "The Digital Humanities Debacle: Computational methods repeatedly come up short" by Da (2019b) and reactions to `https://www.chronicle.com/article/The-Digital-Humanities-Debacle/245986` summarized as "Digital Humanities Wars", and the "Special Forum on Responses to Nan Z. Da" in *Cultural Analytics* on `https://culturalanalytics.org/section/1580-debates`

[2] Nan Z. Da (2019b) could be summarized in this respect as suggesting that no more funding should be invested or wasted in computational literary studies.

Developed against the background of (literary) analysis of literary texts, it should also—with slight adjustments if necessary—be generally suitable for text analysis. Because the complexity analysis focuses on the project's methodological approach, whereby the dimensions proposed are independent of the school of thought or interpretational theories behind them, the model is thus suitable for all text analysis procedures: those understood as belonging to a structuralist tradition, as well as those that implement more postmodern or other approaches. The model maps five dimensions that are relevant for any computational text analysis approach; these are oriented toward essential aspects of research approaches that use computational text analysis, i.e. towards (i) the phenomena of interest, (ii) the texts being studied, and (iii) the insights gained in the process.

The following paper presents this model by first introducing the superordinate aspect, then explaining associated complexity dimension(s), elucidated through examples from my own current research and three other literary approaches already published as pamphlets in this series. Finally, I summarize the model once again and outline the possibilities for its general application.

2 The complexity of phenomena

Literary text analyses usually focus on phenomena, in the sense of certain characteristics of texts. In literary studies, the notion of phenomena is difficult to grasp, and the attempt to grasp certain phenomena on the basis of concepts is subject to extensive theoretical debate. In the field of computational text analysis, the operationalization required provides a suitable basis for considering phenomena of various kinds in terms of their complexity. This is due to the fact that concepts used to capture the phenomena in question must be operationalized in order to be computationally implementable. The handling and reflection of operationalization is exemplary for the development of the Digital Humanities towards reflected approaches in the last decade. At least since Franco Moretti's recent contribution "'Operationalizing': or, the Function of Measurement in Literary Theory," this procedure has become an important catchword in summarizing the field's self-image, as "operationalizing has certainly changed, and radicalized, our relationship to concepts" (Moretti 2013, p. 119).

In the proposed model, the operationalization of the phenomena is used to assess two complexity dimensions of the composition of concepts of phenomena and their contextualization. It should be noted that the focus is not on whether a definition of a certain phenomenon is as generally valid as possible, but rather on the description of the phenomenon used by the researchers, or what can be deduced, from their approach, about the operationalization of the phenomena.

The description or operationalization of the same phenomenon can therefore vary significantly in different research projects. The consideration of the complexity associated with phenomena or their operationalization is a procedure in line with Alvarado's demand: "As humanists, we should not accept the glib premise that the most easily operationalized ideas are the best ideas, but should instead engage in an overt and critical review of operationalization as a form of argument, even as we employ this form to test and explore a grand theory" (Alvarado 2019).

2.1 Complexity dimension 1: The composition of phenomena

The first dimension for which the complexity of computational text analysis can be determined is the composition of the phenomenon under investigation. The question is: Is the phenomenon considered simple, not further subdividable, or is it composed of several phenomena? An ongoing research project on gender and illness in literary prose texts[3] deals, for example, with the phenomenon of illness. This phenomenon can be interpreted in fundamentally different ways: One could, for instance, determine the illness of a literary character by whether it is treated by a doctor, but one could also use a series of phenomena such as physical reactions, statements of the character, etc., to determine an illness. The former would be an example of a simple phenomenon, the latter an example of a compound phenomenon. As another example, let us consider the composition of phenomena or their operationalization in concrete research projects. The overarching question in the project just mentioned—the possible connection between illness and the gender of a character—is operationalized as a composite phenomenon. Illness effects are examined on the basis of characterizations of figures, the latter being determined on the basis of a series of phenomena.

The approach described by Weitin & Herget (2016), on the other hand, deals with a simpler phenomenon. Among other things, the question of whether there are so-called "Falcon Topics," i.e. topics that can be seen primarily as representations of topics or summaries of the plot of specific texts, is addressed by considering topics determined by topic modeling. Although the topic-modeling process itself is complex,[4] the phenomenon captured by the so-called Falcon Topics is simple. It consists of a series of words that co-occur more or less frequently. Weitin's contribution (2018) deals with the question of the formal design of novellas and their "average" form. This is translated into the question

[3] Cf. e.g. Gius et al. (2019), Adelmann et al. (2019) and `https://www.herma.uni-hamburg.de/subprojects.html`.

[4] For the complexity of or the data generated by computational processes, see explanations in the below section "The complexity of gaining insights."

of the distinctiveness of a novella within a series of texts, which is addressed by modeling network measures that use the results of stylometric analyses. In this way, a literary historical question is translated into comparatively simple phenomena: the relative frequency of words in the context of a group of texts. Finally, Krautter et al. (2018) focuses on the question of how a character typology in drama, especially in the form of a distinction between protagonists and other characters, can be made. This typology is based on aspects such as the amount of speech, themes, and interaction of the characters; it models character types accordingly as a composite phenomenon.

One can already see from these brief descriptions that within an approach, several phenomena are often considered, which are combined into one superordinate phenomenon. Those phenomena that make up a more complex phenomenon can in turn be considered individually in terms of their complexity. For example, when the Gender and Illness project is concerned with analysis of the effects of illness on characters, the characterization of the figures mentioned above is presented as a complex phenomenon composed of character expressions and character description. These two phenomena, in turn, are used as simple phenomena in the approach and are not differentiated further according to types of utterances or character actions, character traits, etc. For the assessment of complexity, in the sense of the composition of a phenomenon, one can therefore use not only the composition but also the number of analyzed phenomena as a criterion.

Even if the composition of phenomena is in part very dependent on the concrete research approach, there are phenomena which have a more generally valid specific composition. For example, in the above, this applies to the concept of topic in topic modeling. Phenomena relevant to many approaches, such as co-reference, are usually operationalized in a more general way. Basically, this defined complexity in the sense of the composition of phenomena applies in particular to phenomena which are analyzed in Natural Language Processing (NLP) and for which a computationally implemented operationalization is already available. When phenomena are operationalized in NLP procedures with recourse to established concepts, it is therefore usually comparatively easy to identify the individual phenomena of a compound phenomenon. In the case of co-reference, these phenomena are entities, references, etc., which are determined in a relatively general way.

In contrast, descriptions of phenomena or operationalizations that are not based on concepts that are already used in computational procedures are often more difficult to determine, nor is the operationalization of a phenomenon always explicit. Nevertheless, up to a certain point it can usually be reconstructed from publications relevant for a specific approach. As a general rule, the complexity of a project, in terms of the composition of phenomena alone, does not provide any indication of the

relevance or quality of the approach. Phenomena that are operationalized as particularly complex in their composition are in themselves no better or worse than simpler ones.

2.2 Complexity dimension 2: The contextualization of phenomena

In addition to determining the composition of phenomena, it is also a question of what kind of knowledge is used for their identification according to the operationalization. Accordingly, the core question is: Do we need knowledge beyond textual knowledge to identify a phenomenon? Since the complexity model was developed for computational text analysis, it is reasonable to assume that the analyzed text itself always plays a role in the identification of phenomena in this text. Correspondingly, the question is whether beyond the information available in the text,[5] further knowledge, such as special domain knowledge, additional (intra-fictional or extra-fictional) world knowledge, etc., is used. This additional knowledge can be in the form of lexicons, databases, but also other texts, etc. In the examples discussed, the analyzed phenomena are mainly developed by referring to the analyzed texts. Knowledge beyond this is only used in the computational analyses to the extent that it serves to evaluate the results of the computational analysis. In the approaches to gender and illness and to character types, knowledge from secondary literature in literary studies is used to generate (evaluation) data or to evaluate the results. Character representations in the course of the text or the classification of character types are compared with this specialist knowledge.

The fact that all the approaches discussed here relate their results, in one way or another, to findings in literary studies is irrelevant for the assessment of the complexity dimension of contextualization. In particular, the contributions on the Falcon Topics and the network analysis of novellas take a position rooted in literary studies that goes beyond pure textual reference and makes literary-historical or methodological considerations. However, this takes place on a superordinate level of reflection beyond the computational method. Accordingly, there is also no compelling connection between complexity in the sense of contextualizing phenomena and the literary-historical characteristics or connectivity of an approach.

In any case, the following also applies here: The classification of complexity applies to the use case under consideration. Other cases may have different degrees of complexity for the same phenomena and, furthermore, these cannot be directly translated into an assessment of the quality of an approach.

[5] The complexity of the phenomenon could also be further differentiated according to the extent of the intratextual context needed for the determination or the textual extent of the phenomenon itself (see also my remarks in Gius 2016:11–15). This and other further differentiations are not made here in order to keep the model as simple as possible for the time being. Further differentiations are to be made in the application of the model as needed.

Notwithstanding this, advocates of certain methods of literary studies have a clear preference here for recourse to additional knowledge and specific types of additional knowledge. For example, approaches in the post-structuralist tradition can be distinguished from those in the structuralist tradition, among other things, to the extent that the former include considerably more contexts in the observation of phenomena.

Figure 2.1 shows the two dimensions that determine the complexity of the phenomena investigated or their operationalization for the phenomena discussed. There you can see, among other things, that in the project Gender and Illness a comparatively simple concept of illness is used (illness is analyzed via word fields), while the gender role problem is negotiated in a somewhat more complex way (it is considered via figure characterization in the sense of a complex phenomenon). It can also be seen that figure interaction as a phenomenon that is part of the character types is correspondingly less complex than these.

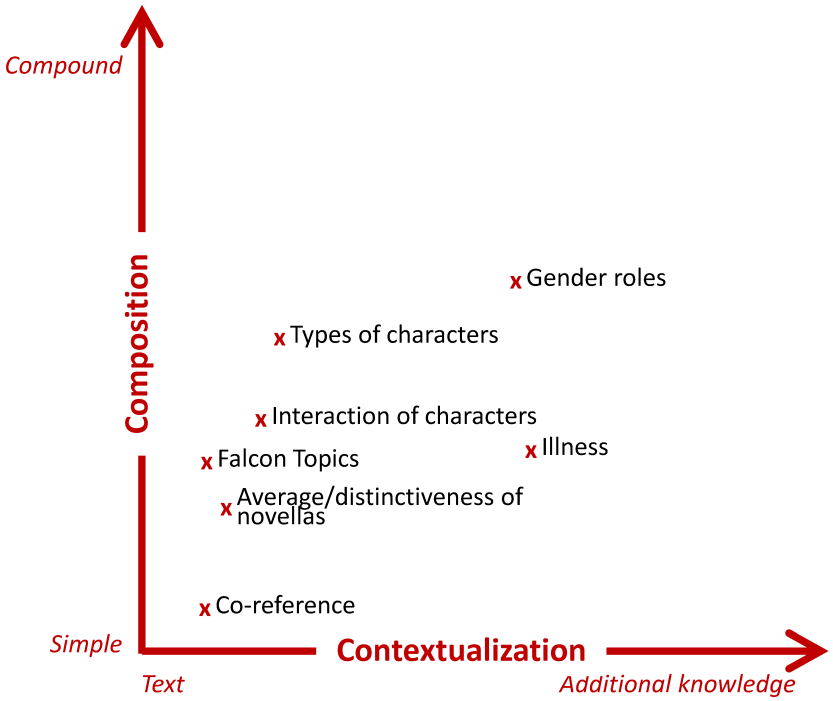

Figure 2.1: Complexity dimensions for phenomena

3 The complexity of texts

In the course of digitalization and Big Data, texts are the focus of attention in a new way. The supposed promise of the Digital Humanities is that a new era has now dawned in which we can analyze masses of texts that we have not or simply cannot read. In fact, however, the field of computational analysis of literary texts seems to be divided in at least two positions in relation to the amount of text analyzed. There are indeed approaches that analyze a very large number of texts. However, these are mostly computer-driven approaches in which philological quality criteria play no significant role either in the selection or in the analysis. Approaches such as the one of Michel et al. (2011), which became famous under the keyword "Culturomics," therefore tend not to yield any findings relevant for literary studies. From a disciplinary point of view, approaches in which the authors draw on literary expertise and in some cases, profound knowledge of the text, and those in which correspondingly less-extensive text holdings are considered tend to be much more interesting.[6]

3.1 Complexity dimension 3: Text heterogeneity

From the point of view of computational text analysis, the question whether or not one deals with (supposedly) Big Data is therefore mainly interesting insofar as it is connected with the question of whether or not one knows the texts one is analyzing. With regard to the complexity of the texts used, however, the more comprehensive question is more relevant: How many (how) different texts are analyzed? The heterogeneity of texts is determined by the number of texts themselves, but also by the number of different text properties that are or could be relevant to the question. In the case of literary texts, these typically include characteristics such as literary form, genre, epoch, author's gender, place of publication, etc.

With regard to the complexity of a project, this also involves the question of how many texts and text properties play a role, whereby the scale ranges from one text to very many, very heterogeneous texts.[7]

A comparatively high degree of text heterogeneity exists in the project Gender and Illness. The corpus consists of over 2,000 German-language texts of one main genre but many sub-genres, written by various authors from different epochs between naturalism and modernism. In the Falcon Topics and the network analysis, 86 German texts from the *Deutscher Novellenschatz* are used. This means that

[6] Cf. also Herget & Weitin (2016, p.4).

[7] This is not a single scale, strictly speaking, because there are two pairs of opposites that are mapped: number (of texts) and heterogeneity (of text properties). However, these properties are combined into one dimension because they increase the complexity of texts comparably. Comparable simplifications were also made for the other dimensions.

the corpus of a genre (in this case, novellas) of 82 authors covers several epochs between classicism and realism. Due to their compilation in anthologies, the texts can be assumed to have a certain homogeneity. Accordingly, the text base is of medium heterogeneity. A corpus of 98 German-language dramas, from several epochs from the Enlightenment to the modern age, and written by various authors, was used for the character typology. Four data sets with 39 to 42 dramas each were created from these, most of which are considered individually or in comparison to another. Thus, the text basis is also of medium text heterogeneity.

Since text heterogeneity can encompass heterogeneity in various respects, the complexity of this dimension can be increased in many ways. The complexity dimension of text heterogeneity for one approach, therefore, can best be determined in relation to other approaches. For example, medium-size corpora are relatively variant in the context of single text analyses and rather less in the context of analyses using thousands of texts. Furthermore, even relatively small amounts of texts may be considered very variant when considering individual text phenomena, if they show a great diversity with respect to the phenomenon under consideration. However, even in the case of text heterogeneity, there is no evaluation of the quality of an approach that goes hand in hand with its complexity, since all forms are potentially interesting from the point of view of literary studies.

Figure 3.2 illustrates the text heterogeneity of the discussed procedures in relation to each other.

4 The complexity of gaining insights

In addition to considering the phenomena and texts used, the assessment of the complexity of computational text analysis also addresses the overarching question of how the computer is used to generate insights. In this context, "insight" refers to the results of a text analysis process. These include an assignment of less-elaborated phenomena to small text segments as well as findings on literary-historical contexts.

When considering the handling of knowledge in computational text analysis in terms of its complexity, two dimensions of complexity should be considered. On one hand, there is the question of the analytical approach: Who analyzes? On the other hand, the question of the integration of the results of computational analysis into the knowledge process: How do the results contribute to insights? Since both dimensions may involve elaborate computational procedures, in some cases artificial intelligence

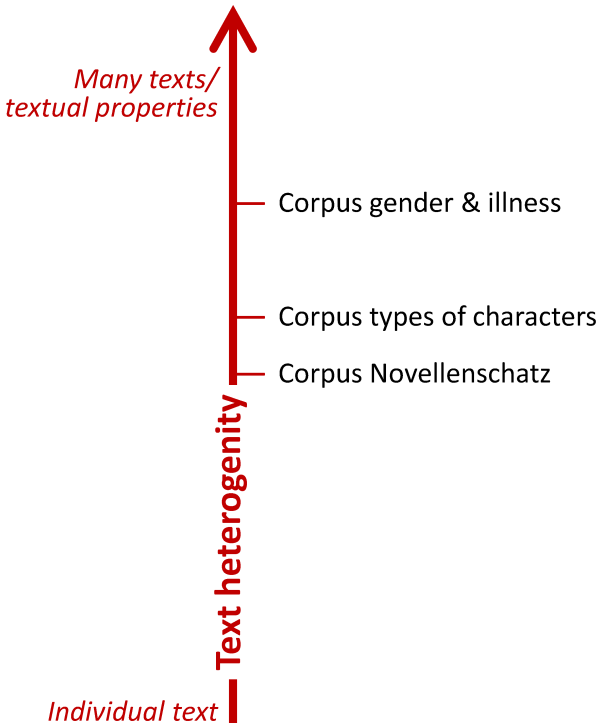

Figure 3.2: Complexity dimension for text(s)

plays a role. On one hand, this means that the complexity of these two dimensions is particularly difficult to determine. On the other hand, these are the aspects that are often particularly controversial in the debate about the relevance of computational text analysis in the humanities.

Yet critics often fail to recognize that informed discussion of these aspects should be a central component of the humanities. One can understand computational text analysis with Luhmann as second-order observations, since they are observations of observations that are analogous to reading.[8] In analogy to the development of professional reading, literary studies—or the humanities in general—therefore have the important task of describing procedures of 'computational knowledge generation' and (further) developing them in the humanities.

4.1 Complexity dimension 4: Mode of analysis

The complexity dimension of the mode of analysis is about who produces the findings. As already explained, findings are all conceivable results of analysis. Therefore, the question is who identifies the (operationalized) phenomena in the text base. The main question is: Is the text base made ac-

[8] This analogy should not imply that the computer reads or even thinks. However, the computational analyses that are carried out can be seen as observational procedures even without this misleading anthropomorphization of the computer.

cessible by humans or by computers? Generally, it is assumed for all approaches that the computer is used in both cases, as this is, after all, a model for computational approaches. In addition, the results of the process of analysis are usually further analyzed by researchers, i.e. by humans. What is interesting about the mode of analysis is mainly how the data for the gaining of insights in an approach is generated, whether by human analysis or by using the computer. Reading in itself is of course a highly complex process. In a computational text analysis approach, though, it is usually taken for granted without further reflection. However, complexity increases with the increased use of the computer, because highly computer-based approaches are more complex in that they are harder to understand for humans (e.g. in the case of deep learning) or very complex to implement when they model a phenomenon in detail (e.g. in rule-based systems). In computational approaches, human reading and text processing typically results in annotations of text passages or at least in the addition of meta-information to the text, whereas machine processing mainly involves text mining, which in turn generates additional data. Both modes of analysis can be further differentiated, for example, according to interpretation theory (e.g. into text-, reader- or author-oriented approaches) or the applied machine method (e.g. into rule-based and learning methods).

In concrete research projects both modes almost always occur. For example, in the Gender and Illness project, manual annotations of text passages and semi-automatic methods for word-field generation for further processing or method development are combined with automatic methods for character recognition, segmentation, and sentiment analysis. Since the intermediate steps in the analysis are mostly checked manually and are in some cases supplemented, this is a procedure between reading and automatic processing and thus of less complexity.

For the recognition of the Falcon Topics, a topic modeling approach with MALLET is chosen. Although this is considerably shaped by the manual revision of the stop words and by experiments with the parameterization, the focus is on automatic processing, which makes the approach comparatively more complex.

The situation is somewhat different with the network analyses of the *Novellenschatz*. There, existing procedures for network generation are used, whose selection and parameterization are motivated by literary studies, but the application is primarily one of computational processing.

A mixture of manual and automatic approaches is also used for character typology. The focus, however, is on automatic approach, since a classification of character types is implemented as a machine-learning procedure that uses the token number of figure speech as well as data from topic modeling and network

measurements as features. Thus, the approach is rather complex.

The activities described here suggest that work steps such as preprocessing, feature identification and parameter manipulation are aspects that can be introduced as additional criteria when considering the complexity dimension mode of analysis. After all, these are sometimes very elaborate manual or computational data-preparation steps, which are not only usually very time-consuming, but can also have a considerable impact on the result of the analysis.

4.2 Complexity dimension 5: The production of insights by computational analyses

In literary studies practice, the most complex task of text analysis in terms of gaining insights is that of interpreting the textual basis as a whole with regard to the chosen question. However, text interpretation in the sense of literary studies is not the focus of computational approaches to literary texts, and the idea of the computer as an interpreter of literary texts appears to be hardly plausible so far. Nevertheless, it is worth thinking of computational interpretation as an extreme of the dimension of gaining insights. Following the practice of literary studies, the complexity dimension of the contribution of computational analyses to the insights and findings can be seen as covering all processes from the first analysis of the text focusing on text comprehension to the interpretation of the text basis as a whole.[9] Alternatively, following Peirce (1935), the categorization of research logics in deduction, induction, and abduction, which is particularly widespread in the social sciences, can be used as a scale for describing the epistemological contribution of computational analyses.[10]

Regardless of the question of which systematics one uses for activities concerned with understanding texts, the central question in the last complexity dimension is: How far does the contribution to insights and findings by computational methods go? The possible contributions to findings range from simple text analysis to interpretation, or from deduction to induction or abduction. Roughly, the complexity levels can be described as follows: If already formalized analysis categories and procedures are used to analyze texts, the analysis is deductive. If the analysis aims to develop new categories of analysis or to reveal certain relationships, it is more of an inductive procedure. Finally, if the objective is to develop hypotheses about larger, newly discovered connections in the texts, it is abduction or interpretation.[11] The first two activities, deduction and induction or text analysis in the narrower

[9] Cf. Winko (2003).

[10] Cf. also the work in the project hermA, in which various logics of research were examined in the context of annotations (among others Gaidys et al. 2017 and `www.herma.uni-hamburg.de`).

[11] Cf. also Eco (1987): "[D]er Text ist ein Objekt, das die Interpretation im Verlauf ihrer zirkulären Anstrengungen um die eigene Schlüssigkeit bildet auf der Basis dessen, was sie als ihr Resultat erschafft. Ich schäme mich nicht, da ich auf diese Weise den alten und immer noch gültigen hermeneutischen Zirkel definiere. Die

sense,[12] rather correspond to what Frege (1982) defines as sense, while the abductive or interpretative mode deals with Frege's concept of textual reference. In any case, the question of the contribution to insights and findings is centered on the assessment of the novelty of the computationally generated results, i.e. the innovative findings they contain, before possible further analysis and interpretation.
When dealing with the complexity dimension of the contribution to insights, it should be noted that in a typical literary text analysis, all modes are usually present and blend seamlessly into one another. As mentioned above, it is interesting for the complexity assessment which modes are to be supported computationally; thereby, it should be comprehensible to what extent the mode satisfies the principles of literary studies. Accordingly, deep learning procedures, which certainly have analogies to inductive—if not abductive—procedures, are not considered abductive or interpretative in the complexity model as long as they have not been developed according to criteria of literary studies or can be assessed according to these criteria. In principle, however, one can assume that rule-based procedures are less complex than machine-learning procedures, whereby here again the supervised procedures are less complex than the unsupervised ones.[13] However, a consideration of the general dependence between computational methods and the complexity of their contribution to knowledge is still pending.

Due to the mostly mixed modes, it is particularly important in this dimension to differentiate accordingly in the consideration of an approach. In the case of the project on gender and illness, for example, the change in the constellation of characters will be analyzed deductively on the basis of the character mentions. An inductive procedure is present when gender categories are worked out by clustering character names (which are then used deductively in the analysis as a new procedure). And finally, an abductive approach is present when a new element influencing character illness is discovered through an overall observation.
All in all, in this example, the approach is rather deductive. Mainly simple extraction methods are used, which partly aim at the development of a computational method that automates character analysis and accordingly works more inductively or more demandingly on a text-analytical level.
In the contribution by Weitin & Herget (2016), the aim is to examine topic modeling for its suitability for literary text description. The underlying idea that topics may be suitable for describing individual

Logik der Interpretation ist die Peircesche Logik der ‚Abduktion'" (p. 45).

[12] "Text analysis" is literary-scientifically ambiguous, since the term means both a text analysis that focuses on the understanding of the text and serves as precondition for the subsequent interpretation, and the process of analysis and interpretation as a whole, see Winko (2003).

[13] For a description of the use of computational procedures in computational text analysis see Underwood (2015), who proposes seven types.

texts is a literary-studies view on the applicability of the method. Correspondingly, the purely computationally possible gain in insights has to be estimated higher here. In the context of the Falcon Topics, topic modeling can at least be regarded as an inductive method, which is located between text analysis and text interpretation.

The situation is similar with network analyses, in which the suitability of the computational methods for literary studies is also examined. Since the procedures are found to be suitable for the analysis of the novellas, an automatic procedure is also available here, from which the gain in insights is comparatively high.

The character typologies, in turn, use a number of less complex (deductive) procedures to extract features, which they then transfer into a classification, which is a more insight-generating process that can accordingly be described as inductive. Figure 4.3 illustrates the discussed generation of findings and insights of the projects exemplarily on the corresponding complexity dimension.

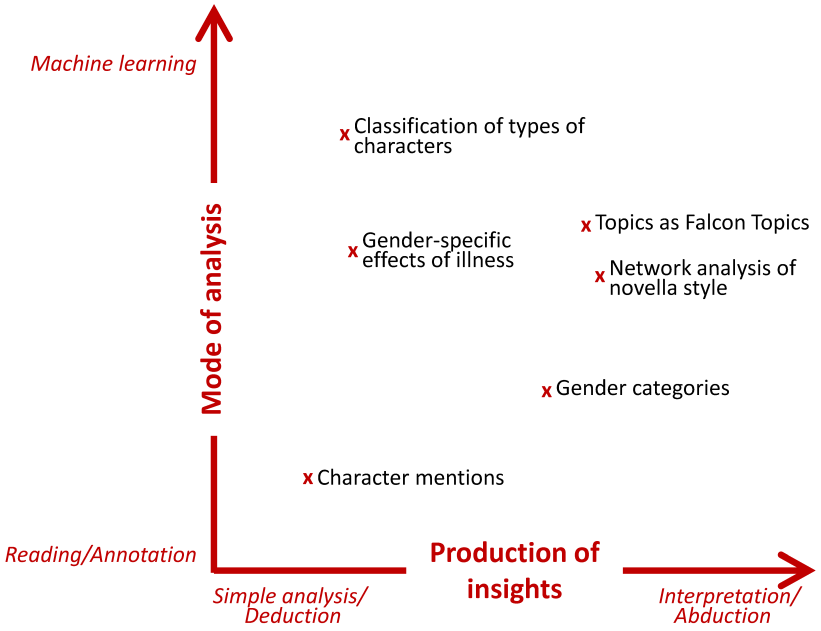

Figure 4.3: Complexity dimension for gaining insights

5 The model for the complexity of computational text analysis

The presented model is designed as a general tool which can be used for the consideration of computational text analysis. The complexity aspects should be specified or adapted according to the application

situation. The possibilities of application have already been sketched exemplarily. Figure 5.4 shows the complexity of the discussed projects in all five dimensions, whereby the discussed aspects were summarized to an overall assessment of the individual dimension for each project.

Once again, please note the simplifications made in the design and exemplary application of the model. For example, contextual knowledge was not differentiated further, the number of texts and text properties were subsumed under heterogeneity, no distinction was made between literary analysis and interpretation, and their connection with Peirce's research logic and computational procedures was not examined in detail. At some points in the description and application of the model, possible further criteria that can be incorporated into the model have already been highlighted, such as parameter manipulation and preprocessing in the fourth complexity dimension, the mode of analysis.
On the whole, the model is to be understood as a first draft, which now has to prove itself in its application, whereby possible extensions are expected to become apparent.

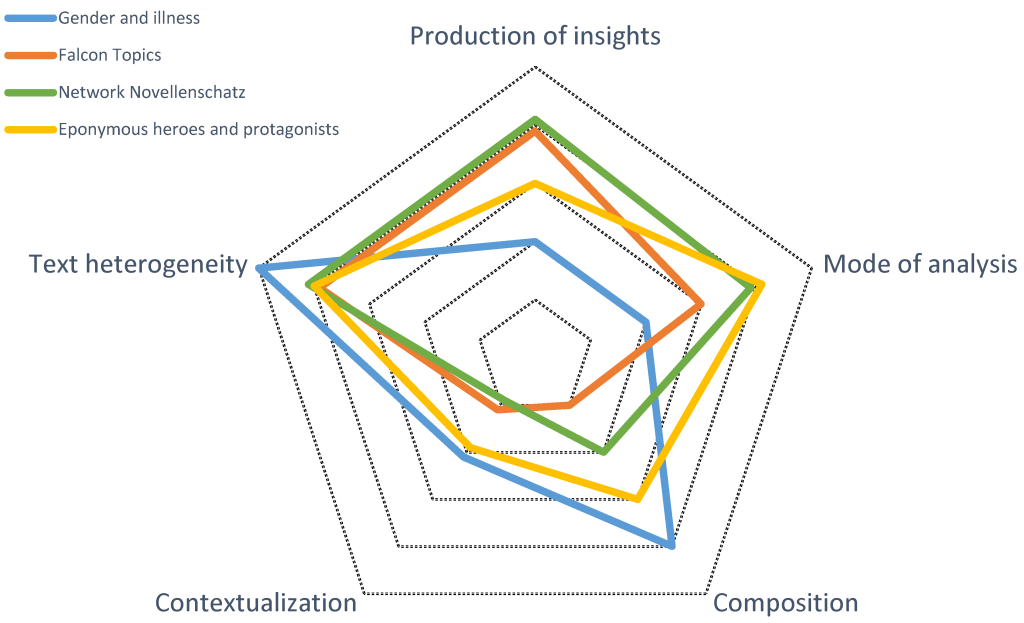

Figure 5.4: Complexity of the discussed approaches in all five dimensions.

The proposed model is presented below in a condensed form. It is important to note that it can be used for the representation of partial activities, as well as for the representation of an entire research project (cf. figures 2.1, 3.2, 4.3 and figure 5.4, respectively). Furthermore, it is possible to consider only a selection of the five dimensions.

- **Complexity dimension 1: The composition of phenomena**

  Question: *Is the phenomenon considered to be simple, not further subdivided, or composed of several phenomena?*

  Complexity: from simple to complex phenomena

- **Complexity dimension 2: The contextualization of phenomena**

  Question: *Does one need further knowledge beyond textual knowledge to identify a phenomenon?*

  Complexity: from textual knowledge to various types of extensive additional knowledge

- **Complexity dimension 3: Text heterogeneity**

  Question: *How many (how) different texts are analyzed?*

  Complexity: from a text with homogeneous characteristics to many texts that are heterogeneous in themselves and in relation to each other

- **Complexity dimension 4: Mode of analysis**

  Question: *Is the text base analyzed by humans or by computers?*

  Complexity: from annotated by humans to analyzed by machines through learning

- **Complexity dimension 5: The production of insights by computational analyses**

  Question: *To what extend does the computational method contribute to gaining insights and findings?*

  Complexity: from the application of simple rules to individual text elements to the interpretation of the entire text base

6 On the application of the model

As already explained, the determination of the complexity of a computational text analysis approach primarily concerns the normative decisions by the researchers in the five dimensions. The decisive factor is not so much how phenomena, texts, and gain of insights and findings as such should be modeled, but rather how they are actually implemented. A productive critique of an approach should also be based on this. If, for example, one looks at an approach that models a phenomenon in a highly simplified way in all five dimensions, this is more productive than if one limits oneself to criticizing the under-complex representation of the phenomenon and thus ignoring any further aspects of the

approach. In particular, it is not very illuminating to discuss a project that, for certain reasons, uses a very simple operationalization of the phenomena mainly in relation to this operationalization. Or when a lack of literary historical knowledge is called out, when work is done on the recognition of certain phenomena on the basis of selected texts that are only partially motivated by literary studies. For, even if in some cases a less-complex implementation in one dimension is sub-complex in the sense of not being adequate to the subject, it is often not the only reason for the issues raised by an approach. If one considers the interconnections between the individual dimensions in the complexity model, one sees that the accusation of undercomplexity is only valid to a limited extent in many cases. This is because increasing the complexity in one dimension nearly always results in increased overall complexity. If, for example, as a convinced post-structuralist, I would rely not only on the texts examined but on further contextual knowledge for the recognition of phenomena, and, furthermore, want to access my corpus computationally, the latter becomes correspondingly more demanding. The same applies if I would want to examine numerous, very heterogeneous texts in relation to a phenomenon and have to operationalize it accordingly in view of its heterogeneity. Or if I want to implement interpretation or at least abductive processes computationally, then even the simplest operationalization in a few uniform texts becomes a great challenge. It is important to bear this dynamic in mind when demanding that certain aspects of an approach be implemented in a more complex way in a positive sense. In addition, one should also consider that the more complex an approach is, in the sense of the model presented, the more numerous the possible errors would be—and the more demanding both the criticism and the implementation of the approach.

This pragmatic aspect of complexity, especially, makes the model a planning tool beyond its function as a heuristic for the structured critique of one or more approaches. It can also be used as an instrument for designing an approach, and it should be used in all phases of computational text analysis: from the design of the research approach at the beginning of the research work, through repeated assessment or readjustment during the course of the project, to the evaluation of the results at the end and the reflection of the entire process.
In addition to project-specific questions that can be addressed, the characteristics and combination of the five dimensions can be used for the following purposes:

- Assessment of innovation and risk: What is particularly complex requires particularly thorough consideration. This is where the greatest innovation potential lies—and the greatest risk. Ideally,

the assessment is carried out both for individual tasks and for the entire project, in order to obtain an additional assessment of the range of complexities.

- Resource planning: The less pronounced the individual dimensions are, the faster the project can be processed. Conversely, a tendency towards high complexity means a greater workload. Smaller projects (for smaller contributions, project work, etc.) should tend to be located inwards in the complexity model or focus on inwardly located aspects. In contrast, projects with more substantial time and personnel resources should be located outwards.

- Readjustment: If projects turn out to be more complex as they progress, simplifying one dimension can significantly reduce the overall complexity and thus the resources required. Conversely, this could also be increased if capacities are available. Both can be used specifically where resources are lacking or available. This applies to projects by individual researchers, as well as to team projects or joint research projects.

7 Bibliography

Adelmann, Benedikt, Melanie Andresen, Anke Begerow, Lina Franken, Evelyn Gius, and Michael Vauth. 2019. „Evaluation of a Semantic Field-Based Approach to Identifying Text Sections about Specific Topics“.*DH2019 Book of Abstracts.* Utrecht. `https://dev.clariah.nl/files/dh2019/boa/0895.html`.

Alvarado, Rafael C. 2019. „Digital Humanities and the Great Project: Why We Should Operationalize Everything and Study Those Who Are Doing So Now“. In *Debates in the Digital Humanities 2019*, Ed. Matthew K. Gold and Lauren F. Klein. University of Minnesota Press.`https://doi.org/10.5749/j.ctvg251hk`.

Bode, Katherine. 2020. „Why you can't model away bias“. Preprint: *Modern Language Quarterly* 80 (3). `https://katherinebode.files.wordpress.com/2019/08/mlq2019_preprintbode_why.pdf`.

Da, Nan Z. 2019a. „The Computational Case against Computational Literary Studies“. *Critical Inquiry* 45 (3): 601–39. `https://doi.org/10.1086/702594`.

Da, Nan Z. 2019b. „The Digital Humanities Debacle“. *The Chronicle of Higher Education*, 27. March 2019. `https://www.chronicle.com/article/The-Digital-Humanities-Debacle/245986`.

Eco, Umberto. 1987. *Streit der Interpretationen.* Ed. Rolf Eichler. Konstanzer Bibliothek 8. Konstanz: Universitäts-Verlag.

Frege, Gottlob. 1892. „Über Sinn und Bedeutung“. *Zeitschrift für Philosophie und philosophische Kritik*, 100 (1), 25–50.

Gaidys, Uta, Evelyn Gius, Margarete Jarchow, Gertraud Koch, Wolfgang Menzel, Dominik Orth, and Heike Zinsmeister. 2017. „Project description – hermA: Automated modelling of hermeneutic processes“. *Hamburger Journal für Kulturanthropologie.* `https://journals.sub.uni-hamburg.de/hjk/article/view/1213`.

Gius, Evelyn. 2016. „Narration and Escalation. An Empirical Study of Conflict Narratives“. *Diegesis* 5 (1): 4–25.

Gius, Evelyn, Katharina Krüger, and Carla Sökefeld. 2019. „Korpuserstellung als literaturwissenschaftliche Aufgabe“. *DHd 2019 Digital Humanities: multimedial & multimodal Konferenzabstracts*, Frankfurt & Mainz: 164–166.

Krautter, Benjamin, Janis Pagel, Nils Reiter and Marcus Willand. 2018. „Titelhelden und Protagonisten - Interpretierbare Figurenklassifikation in deutschsprachigen Dramen“. Ed. Thomas Weitin. *LitLab Pamphlets*, Nr. 7. `https://www.digitalhumanitiescooperation.de/wp-content/uploads/2018/12/p07_krautter_et_al.pdf`.

Michel, Jean-Baptiste, Yuan K. Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, J. P. Pickett, i.a. 2011. „Quantitative Analysis of Culture Using Millions of Digitized Books". *Science* 331 (6014): 176–82. https://doi.org/10.1126/science.1199644.

Moretti, Franco. 2013. „'Operationalizing': or, the Function of Measurement in Literary Theory". *New Left Review*, Nr. 84 (December): 103–19.

Peirce, Charles S. 1935. *Collected Papers of Charles Sanders Peirce, Volumes V and VI: Pragmatism and Pragmaticism and Scientific Metaphysics.* Ed. Charles Hartshorne and Paul Weiss. Cambridge, Mass: Belknap Press of Harvard Univ. Press.

Div. authors. „Special Forum on Responses to Nan Z. Da". 2019. Journal of Cultural Analytics. 17. September 2019. https://culturalanalytics.org/2019/09/special-forum-on-responses-to-nan-z-da/.

Weitin, Thomas. 2018. „Average and Distinction. The Deutsche Novellenschatz Between Literary History and Corpus Analysis". Ed. Thomas Weitin. *LitLab Pamphlets*, Nr. 6. http://bit.ly/2CBXNol.

Weitin, Thomas, and Katharina Herget. 2016. „Falkentopics". Ed. Thomas Weitin. *LitLab Pamphlets*, Nr. 4. https://www.digitalhumanitiescooperation.de/pamphlete/pamphlet-4-falkentopics/.

Winko, Simone. 2003. „Textanalyse". In *Reallexikon der deutschen Literaturwissenschaft: Neubearbeitung des Reallexikons der deutschen Literaturgeschichte*, Ed. Harald Fricke, Klaus Grubmüller, Jan-Dirk Müller, und Klaus Weimar, 3., neubearb. Aufl. Berlin: De Gruyter: 597–601.