The Measurement of Text Quality

Eivor Finset Spilling

Abstract: This article discusses a type of study that is based on the naturalistic view of science, but where the object of inquiry is texts – man-made products of meaning. A specific study of texts written by beginning writers is used as a starting point for discussion. This study applies quantitative methods and measures meaningful structures and the quality of the texts through objective and systematic inquiry. This contrasts with the view of the interpretivist tradition, usually related to a more qualitative research approach, that emphasizes interpretation of texts. The following question is explored: How can a quantitative analysis of text quality handle meaningful structures in text? The article discusses the role of language and the researcher in the making of the analytic categories and in the coding of the texts. Further, the article highlights both the necessity of interpretation and understanding through language, and the procedures offered by quantitative methods to address this.

Keywords: interpretivism, naturalism, meaning, measurement, text quality

Introduction[1]

Texts are important objects of study in different research fields. Underlying assumptions about science will guide how texts are handled and how they are understood as an object of study. Two main approaches can be distinguished: In an interpretivist tradition the researcher typically seeks to understand and interpret the potential meaning in the text through a dialogic process (Gadamer, 2004). In a naturalist tradition, on the other hand, where aims and methods build on ideals from the natural sciences,

1 I would like to thank Vibeke Rønneberg, Wenke Mork Rogne and an anonymous reviewer for reading and commenting an earlier version of the article.

Citation of this chapter: Spilling, E. F. (2021). The Measurement of Text Quality. In T. A. Haugen, S.-A. Myklebost, S. J. Helset, & E. Brunstad (Eds.), *Språk, tekst og medvit* (pp. 47–66). Cappelen Damm Akademisk. https://doi.org/10.23865/noasp.146

text analysis will typically entail some kind of quantification of text features. In a quantitative study of texts, the combination of an object of inquiry that is a man-made product of meaning, and methods originally designed for empirical and objective investigation, can lead to a tension that is worth reflection. The aim of the article is to discuss this tension in relation to a text material where it is particularly pronounced, namely in a study of the texts of young writers.

As a point of departure for the discussion, a specific study of text quality of beginning writers' stories (Spilling et al., 2021) will be discussed. This study applies quantitative methods while investigating the written performance of first graders, both through analysis of different text features and through a score of holistic text quality. The analysis involves both counting of meaningful structures in the texts and judgments of the quality of the texts, and according to a naturalist ideal this should be done objectively. But is it possible to conduct such analyses without any interpretation? This issue is especially relevant when analysing texts written by very young writers, as they do not yet master all writing conventions. In such cases, the analysis can for instance include decisions about whether marks on a paper are meaningful text or scribbles that do not convey meaning. With this as a backdrop, the specific problem to be discussed in this article is: How can a quantitative analysis of text quality handle meaningful structures in text?

The first section of the paper will outline some general features of the naturalist and interpretivist traditions. The proceeding section concerns text quality within writing research, and briefly presents the example study with its measures and how this study relates to the different philosophical schools, specifically through its methods. Then the process of quantifying meaningful structures will be discussed in light of the different views of science represented by naturalism and interpretivism. It will be argued that interpretation and understanding through language is an important foundation for the analysis in studies of text quality influenced by naturalism. Further, it will be demonstrated that rigorous work with explicit coding rules and double rating in analysis of features that to a high degree demand interpretation, is the way that such studies secure transparency, objectivity and replicability.

Philosophical discussions within the social sciences[2]

Naturalism

The view that the social sciences should adopt the ideals from the natural sciences can be named naturalism (Gorton, 2010). The adherents will claim that the social sciences should have the same aims and methods as the natural sciences. This implies that the social sciences should be empirical, seek to find lawlike causal explanations and be value neutral (Gorton, 2010). Being empirical is in this context related to the possibility of testing a theory. The social phenomena studied must be operationalized in a way that makes them possible to measure. This can often imply that phenomena are reduced to smaller parts, like different variables that constitute hypotheses and research questions (Creswell & Creswell, 2018). Observations and experiments can then be conducted to test hypotheses, which again can contribute to the construction and testing of theories. From a naturalist view, theories should give causal explanations of social phenomena. These explanations should be as general as possible, ideally lawlike generalizations, explaining different kinds of phenomena. Value neutrality refers to the claim that scientific evidence cannot imply moral evaluations (Gorton, 2010). Further, external values like the interest of the researcher should not influence the data analysis and the testing of hypotheses and theories. This reflects the view that the object of study can be investigated objectively.

Philosophical worldviews, like naturalism, represent some broad philosophical assumptions that guide the practice of research (Creswell & Creswell, 2018). Philosophical worldviews are interconnected with research designs and research methods, and one way of describing this relation is that these three components inform the overall research approach, which can be qualitative, quantitative or mixed methods (Creswell & Creswell, 2018). Frequently, the philosophical foundation will imply a specific research approach (Creswell & Creswell, 2018). When researchers bring with them naturalist assumptions about the

2 For practical reasons the term "social sciences" is used. This also includes what often is called the humanities.

world, they also tend to apply quantitative designs and quantitative research methods. Typical quantitative designs are experiments and surveys. Quantitative methods for collecting data usually imply specification of the information to be collected in advance, with for instance predetermined and closed-ended questioning, and quantitative analysis and interpretation are typically statistical (Creswell & Creswell, 2018). Different designs and methods are often related to a specific philosophical worldview because the philosophy postulates what warrantable knowledge is, and some methods are often viewed as more suitable than others to obtain this kind of knowledge (Bryman, 1984). According to a naturalist philosophy, knowledge should be objective and replicable, and it should concern the relationship among variables, and generally, quantitative methods are appropriate to provide this kind of knowledge.

Interpretivism

A completely different view of the social world and the social sciences than the one advocated by naturalism, is represented by interpretivism. This tradition points out that the social world consists of phenomena related to human beings, and that these phenomena always carry meaning (Gorton, 2010). According to this tradition, the social sciences should aim at understanding the meanings surrounding us and not strive for making causal explanations. When investigating the social world, attention should be directed to human actions – and to intentions and beliefs underlying these actions and the context out of which these actions arise (Gorton, 2010).

Interpretivism is related to the philosophical movement of phenomenology, which is concerned with how phenomena appear to us – how objects, activities and events appear to consciousness (Moran, 2002). To investigate how the world manifests itself to us is only possible through ourselves, through the first-person point of view. This contrasts with the naturalist ideal of an objective third-person perspective. The phenomenological tradition does not reject the existence of an objective world, but argues that our experience of what exists in the natural world is not an exact copy of this (Moran, 2002). Objects from the natural world are

not seen as isolated objective elements. On the contrary, they are perceived through our consciousness, through intentionality, a directedness towards the objects (Moran, 2002). For instance, when reading a book, the reader does not experience the book from the outside. Rather, the reader intuitively knows how to handle the book as an object and directs her attention to its content. When reading, the reader experiences both the book object and the meaning of the text through her perception and cognitive abilities. Thus, in the phenomenological tradition there is not a strict division between subjects and objects as we find it in naturalism.

Another important philosophical tradition within interpretivism is hermeneutics, which is concerned with interpretation and understanding. Gadamer (2004) seeks to clarify the conditions that enable human understanding, and he uses texts as a point of departure for his theory presented in *Truth and Method* (*Wahrheit und Methode*) from 1960. Understanding, Gadamer (2004) claims, always happens from a point of view – a *horizon*. The horizon of an interpreter is decided by her prejudices, the conscious and unconscious attitudes, beliefs and knowledge that she brings with her. The horizon is not static, rather it is constantly in the process of being formed, and tradition is important in this shaping. Gadamer (2004) writes:

> Every encounter with tradition that takes place within historical consciousness involves the experience of a tension between the text and the present. The hermeneutic task consists in not covering up this tension by attempting a naive assimilation of the two but in consciously bringing it out. (p. 305)

To explore the tension, the interpreter has to project the horizon of the text – try to find out what the text claims and take a stand on these claims. The interpreter should enter into a dialogue with the text where initial prejudices are questioned, and where true prejudices are sought for. Then the horizons of the interpreter and the text can fuse, and understanding about the subject matter may be achieved (Gadamer, 2004). This fusion implies that the horizon of the interpreter has been altered and expanded, and that the interpreter understands better than before.

The hermeneutic circle, which describes the interaction between the parts and the whole, plays an important role in the process of

understanding. In traditional hermeneutics concerned with interpreting and searching for the truth of authoritative, e.g. religious, texts, the hermeneutic circle refers to the relation between the parts and the whole of a text. In Gadamer's theory (2004) the whole is not restricted to a text, but points to the horizon, in which the understanding takes place. He emphasises that "all understanding inevitably involves some prejudice" (Gadamer, 2004, p. 272). Our intellectual basis, our horizon, will guide a preliminary understanding of the parts, and the meaning of the parts will again affect the understanding of the whole, and this interaction between the parts and the whole will continue until we experience that they constitute a coherent unity of meaning. Language is also a central part of Gadamer's work on understanding: "[*L*]*anguage is the universal medium in which understanding occurs*" (Gadamer, 2004, p. 390, emphasis in original). It is language that enables the interpreter to experience the unity of meaning, that is, that enables the text to speak in a way that it makes sense to the interpreter. The inner dialogue takes place in language – we need language to investigate texts, as well as all other objects representing human activity.

Interpretivism, incorporating insights from phenomenology and hermeneutics, is a tradition that typically applies qualitative designs and methods. The relation between interpretivism and a qualitative research approach, just as the association between naturalism and quantitative research, is a tendency and not a strict relationship (Bryman, 1984). In studies of social phenomena, observation and interviews are often used, because they can give rich data (Bryman, 1984). Hermeneutic text analysis is one among many approaches to text analysis. In general, qualitative text analysis is more inductive, nonstatistical and exploratory compared to quantitative text analysis (Roberts, 2000).

Quantitative studies of text quality

Cognitive writing research and text quality

Texts are used as object of study within different research fields. One of these fields is writing research, where one important tradition is cognitive writing research. This discipline arose out of cognitive psychology,

and with insights and methods from this field the cognitive processes of writing could be investigated (MacArthur & Graham, 2016). Pioneering works are for example Hayes and Flower's (1980) model of writing as problem-solving and Bereiter and Scardamalia's (1987) work on the development of writing. The general aim of cognitive writing research is to understand writing from a cognitive point of view – to develop models that explain writing performance, learning and development (MacArthur & Graham, 2016). There is a focus on empirical research and on finding general tendencies of writing performance and learning, which clearly can be traced back to a naturalist view of science. Within this tradition, texts are an important source of data, and one way of using them is through systematic text analysis that yields a measure of text quality (also called writing quality). This measure can for instance be used to give information about writing ability, either of single students or specific populations like first-grade students, it can serve as a factor deciding to what extent a writing intervention has succeeded, or it can shed light on product or process characteristics of writing (Grabowski et al., 2014; Van Steendam et al., 2012).

A single approved and established conception of text quality does not exist. According to Van Steendam et al. (2012, p. ix), the measurement of text quality is a neglected issue in many studies of writing research: "[D]efinitions of writing quality may be absent or unclear, and operationalizations of writing quality may suffer from measurement problems." Text quality can be operationalized in many different ways. Holistic scoring of text quality entails that a text receives a single score, e.g. on a six point scale, that reflects the rater's general impression of the quality of the text (Huot, 1990b). Usually there are some benchmark texts or guidelines on which to base the assessment, and typically these focus on the content, like structure and thematic progression, but surface features, like handwriting and spelling, can also be part of it. Another approach to assessing texts is analytic scoring. Here the rater "give[s] scores to individual, identifiable traits, and these scores are tallied to provide the rating for the paper" (Huot, 1990b, p. 238). The traits that are assessed can vary; many studies of texts written by beginning writers concern text length and spelling, while others include content features. The fact that text quality

is understood in different ways can have different causes. The research questions and the genres that are investigated will for instance influence how text quality is operationalized. Contrasting conceptions of text quality can also be a result of the nature of texts – that they are man-made products with complex meaning potential.

An example study of narratives written by first graders

The study of Spilling et al. (2021) is an example of a quantitative study where text quality plays an important role, and it will be used as an example and point of departure for discussion throughout this article. This study is part of the DigiHand project (Gamlem et al., 2020), which investigates beginning writing instruction with and without the use of digital tablets. The study of Spilling et al. (2021) investigates how writing modality affects text quality in stories written by first graders and whether literacy-related skills moderate this potential effect. Texts were collected from eight different schools in the western part of Norway after students had gone to school for three months. 102 first graders wrote two narrative texts each, one by hand and one by keyboard. The two writing tasks consisted of two different picture prompts, one showing a boy about to drop his ice cream on a cat, and one of a girl about to fall down from a tree. The students were, for both tasks, instructed to write a story about what was happening in the picture. The resulting 204 texts were analysed to decide if modality affected the quality of the texts.

The quality measures in this study had to be adapted to capture central parts of texts written by students who are learning to write. Texts written by very young writers are often simple, short and incomplete. Also, on surface level, these texts can be marked by spelling errors and poor handwriting which can make it hard to identify characters and words. The texts analysed in Spilling et al. (2021) were on average 16 words long (*SD* 11 words), and the longest text was 47 words.[3] Two examples are (all

3 Texts shorter than four words were not included in the analysis. The initial sample of the study was 140 students, and of these 38 students did not manage to produce texts of four or more words in both modalities, which gave a final sample of 102 students and 204 texts.

letters standardized to lower-case, but errors of spelling, spacing and punctuation kept in the Norwegian transcriptions):

Example text 1: *en gut står i is såsken en gut jente gut står i kå så får den eine får isn med 6 kule mn auratda så sed nåke så dakulene dat ne fra isen smilte katten*
"a boy stands in the ice kiosk a boy girl boy stand in line then one of them gets the ice with 6 ball but preciselythen so something happened so whenthe balls fell down from the ice the cat smiled"

Example text 2: *iskø t gut mista is pus etis*
"iceline t boy dropped ice cat eatsice"

All texts were assessed for holistic quality – a commonly used quality measure in studies of written composition. This quality measure reflected the overall quality of the texts, and each text received a score from 0 to 5 where 0 reflected low quality and 5 high quality: Example text 1 was given a score of 5, while example text 2 was scored 3. A rubric with general level descriptions of structure, progression of ideas, coherence and vocabulary was used as guidance in the scoring. In addition, the texts were analysed through a text-analytic approach comprising measures of text length, spelling accuracy, space use accuracy, punctuation (correct use of sentence terminators), vocabulary sophistication, syntax (clause construction) and narrative structure (both on global and local level). The scores were not combined in a single sum score; on the contrary each feature was investigated separately. This approach made it possible to identify potential modality effects on specific features of the texts.

It can be discussed how the text features relate to the quality of the texts. While the holistic text quality measure is a judgment of whether the text is perceived as a good or bad story, identifying the different text features is not automatically in itself a judgment of whether the text is of high or low quality. However, which features to include involves some judgment: In this case features that can be regarded as relevant for written storytelling were chosen. Typically, a good story will have few errors on the microlevel, appropriate vocabulary, varied syntactical structures and fulfill the norms of narrative structure. A measure like

spelling accuracy, either as a count of correctly spelled words or as a ratio of correctly and incorrectly spelled words, will clearly be related to quality, as an important convention for written language is to write according to the written standard: Thus, in general, the higher spelling accuracy, the better. A measure of syntax might be related to quality in a different way. In the example study, the syntax measure was based both on the type of clause, main or subordinated clause, and on the presence or absence of syntactical errors. The texts were given a score based on the number of clauses, where subordinated clauses gave more points that main clauses, and where syntactically correct clauses gave more points than clauses with one or more syntactical errors. Subordinated clauses can be useful to express complex relations, like causal relations that are often used in stories, and therefore one might expect the use of such clauses to affect quality. However, it might be that the quality will increase up to a certain score, but that after this threshold value is reached, the quality is not affected, or affected negatively. Text length is perhaps the variable in the example study that intuitively seems less related to quality. However, the number of words produced has been shown to correlate with text quality in texts by primary-grade children (e.g. Berninger et al., 1992; Dockrell et al., 2015; Malvern et al., 2004). To be able to write a story you need to be able to produce a certain number of words, and for beginning writers, who are learning to write, the production of words in itself is probably closer related to quality than for more experienced writers.

The example study is clearly shaped within the tradition of cognitive psychology – and thus naturalism. The writing of the texts was put under careful instructions to make the conditions of writing as similar as possible. The text analysis was standardized through predefined variables. All the texts were assessed according to a manual with formalized coding rules (this is available on Open Science Foundation: https://osf.io/q8z3u/), and the content of the texts was quantified. Further, statistical calculations were done to find general patterns of the writing performance of the students. An important assumption of the study is that there are objective properties of the object of study that can be investigated by the researcher, which is also in line with a naturalistic view of science. It will, however,

be demonstrated below that insights from interpretivism are also valid for this study.

The process of measuring meaningful structures

The necessity of interpreting and understanding texts

Usually, text analysis entails analysis of meaningful structures, even though it might be possible to conduct formal text analysis where only aspects of the text that do not carry meaning are analysed. The analysis of meaningful structures will always imply some qualitative judgment on a fundamental level. Identification of a meaningful structure, to decide if a text feature is what is being searched for, depends on point of view and criteria of classification, and it has to be judged whether a text feature can be placed in a category or not. Further, some meaningful structures will require more interpretation than others. In the example study meaningful structures on different levels of language were analysed: From the micro level with the features spelling, spacing and punctuation via the meso level with vocabulary and syntax to the macro level with measures of narrative structure. The analysis of the features on the micro level requires less interpretation than the features on the meso and the macro level, as the judgments on micro level can be compared to clear norms. In digital texts, text length and spelling can, at least in theory, be analysed automatically with programmes for character count, word count and spellchecking. However, in text written by beginning writers, the counting of such features is more challenging than in texts written by experienced writers. Beginning writers do not necessarily master the correct form of all letters as they for example do not know the standard form of the letter, or motorically are not able to make well-formed letters when handwriting. Further, they do not necessarily segment words with whitespace. Then several questions arise: Is a slightly bent stroke an <i>, or is it just a line? Is an inverted <p> a <p> or a <q>? Is a mark a line or is it a conventional hyphen-sign? Is a mark a full stop-sign, or just a scroll?

Further, on word- and sentence-level, there will be similar questions. Can a letter combination like *pusvhis* "catwthice" be understood as *pus vil ha is* "cat wants to have ice"? Or can we only identify the start and the end of the letter combination as separate words, *pus* "cat" and *is* "ice" so that this is not really a sentence?

At first glance this letter string does not necessarily make sense. However, the task to which this text was a response, shows a picture of a boy dropping his ice on a cat, and this can guide the analysis. When investigating the letter string, relevant words can be found both at the start and at the end of the string. A researcher who is familiar with beginning writers also knows that segmenting words must be learned, and that a usual strategy can be to only write the first letter of a word. Therefore, one way of analysing this letter string is to split it into more parts and interpret the string as a sentence of four words. The knowledge that the researcher brings with her in the interpretation, her horizon of understanding, is not a disadvantage. On the contrary, it makes it possible to make sense of the text. Thus, prior knowledge of the researcher and an analysis of wavering between parts and whole can make apparently meaningless text parts meaningful.

There is an important distinction between physical objects and meaningful objects: Objects from the natural world, and the laws that govern them, exist independently of human beings (Gorton, 2010). Physical objects can be investigated empirically, and in the natural sciences, knowledge often builds on experience and observation. A text can be observed on a superficial level, e.g. as characters on a sheet of paper or a screen. However, the nature of texts is different from other objects typically investigated in the natural sciences. What constitutes text is meaning. Words, phrases and sentences have an expression, a physical appearance, that can be observed. At the same time, what makes these units entities of language, is that the expression is combined with meaning. Haugen (2021) argues that linguistics, the discipline that studies language, has an intuitive basis that precedes the analytic investigation of and theorizing of language. All language users will have an intuitive understanding of whether a text, a language structure, is acceptable, of whether it makes sense or not. Without this intuitive understanding it is, according to

Haugen (2021), impossible to analyse language analytically. The meaning of a text cannot be observed. Rather, it must be understood.

This point is valid for all research that uses texts, also the study discussed in this article. To code a text – to assign specific values of the different variables to the text – the researcher is obliged to understand the text. For instance, when coding the syntax of a text, which in the example study encompassed type of clause (main or subordinated) and whether the clauses were error-free or contained one or more errors, the reader first has to understand the words of the text. Then, a judgment must be made to decide whether the words constitute clauses, and finally these clauses can be compared to the relevant categories of analysis. A requirement for deciding to which categories the clauses belong, is to understand the meaning of the clauses, and this can only be done through the first-person point of view, as pointed out in phenomenological approaches.

As illustrated above, all analysis of meaningful structures in text is bound to involve some interpretation and understanding. The next section concerns how quantitative studies address this through procedures for systematic, explicit and transparent coding and coding by various raters.

Categories, operationalizations and language

In a quantitative study, the measures are of utmost importance. The measures should be properly defined and properly executed, in order to make a precise description of the object of study (Cartwright & Rundhardt, 2014). To secure accurate measurement it is necessary to find explicitly defined categories, and these categories should be defined according to the purpose of the study (Cartwright & Rundhardt, 2014).

In the study of Spilling et al. (2021), where text quality was measured, the first step was to define essential components of text quality for texts written by beginning writers. This entailed both finding relevant text features that could be quantified, and making a rubric for assessing the quality holistically. The texts were stories written by beginning writers with use of different technologies, and literature on what can be expected with regard to the genre, age group and modality could be the point of

departure for finding relevant components. The rubric for holistic quality defined both what aspects the raters should consider in their assessment, and the different quality levels of these aspects, e.g. that story structure should be evaluated, and that a text of high quality has a complete global story structure, while a text of low quality has no traces of story organization.

The text analytic approach goes further in formalizing the text analysis. The holistic rubric guides the raters, but does not define central concepts, like story structure. Also, when giving a text a single score, the contribution of and the interplay between the different components are concealed. The text analytic approach aims at investigating the different features separately. Thus, rigorous work on operationalizations was done.

Some constructs can be operationalized without much controversy, like spelling. In the example study this was measured as the number of correctly spelled words, and correctly spelled words were defined according to the official Norwegian dictionaries *Bokmålsordboka* and *Nynorskordboka* that correspond to the two written standards of Norwegian. As the spelling measure was a judgment of the ability to spell when composing, more clarifications concerning the coding had to be done compared to coding of spelling of single words (dictation). Rules to handle homographs had to be made: In one text about the boy and the ice, the ice fell *på baken*. Isolated *baken* means "the seat, buttocks", but from the context one can assume that the student intended to write *bakken* "the ground". Further, when analysing compositional spelling, questions about how to separate spelling from grammar and segmentation arise: In example text 1, where the boy gets an ice of 6 balls, it actually says *6 kule* "6 ball" in singular. Is this a grammatical error or a spelling error? In the same example text, a compound was divided; *is såsken* [iskiosken] "the ice kiosk", and two simplexes were written as one word; *auratda* [akkurat da] "precicely then". Are these spelling errors, errors of spacing, or both? Also, rules concerning the relation between *Bokmål* and *Nynorsk* had to be made: Is it acceptable to use both *Nynorsk* and *Bokmål* words in the same text (e.g. *en* "one" – *Bokmål* –, and *gut* "boy" – *Nynorsk* – in example text 1)? Different solutions can be justified with regard to these issues, but in a specific study the criteria for coding the variables have to be clear. With clearly formalized definitions the chances are high that the analysis

of a spelling measure can be executed objectively, in the sense that several raters will code the texts in the same way.

Other constructs are more difficult to operationalize than spelling. A construct of narrative structure can be operationalized in several ways. Literature on story structure (e.g. Labov & Waletsky, 1967; Martin & Rose, 2008; Peterson & McCabe, 1983) shows different ways of analysing this: in different stages/phases, in episodes governed by the goals of the protagonists, in syntactic hierarchies etc. The way that the constructs are operationalized should be decided by the researcher. Is it possible to do this in an objective way? The overall theory of the project, and other similar studies, will provide some guidelines. The researcher should also consider the specific writing task and the context. In the study used as example in this article, relevant questions concerning task and context might be: What kind of instructions did the pupils get? What kind of implicit instructions are conveyed through the context of writing in a classroom? With these answers some operationalizations of (global) narrative structure will be more reasonable than others, e.g. the context of school will imply writing a text with an introduction, a main part and a conclusion, and with labels related to the story genre: orientation, complication and resolution.

In a study of texts by beginning writers, the nature of the texts also makes it challenging to make operationalizations. Young writers' texts will often be short and incomplete, so how to handle ellipses and incoherent parts of the narrative structure, has to be described in detail in the coding rules. Is it for instance possible to make an orientation (introduction) of just one word? Example text 2 can illustrate the question: *iskø t gut mista is pus etis* "iceline t boy dropped ice cat eatsice". The answer can be debated, but a specific study needs criteria that enable the researcher to code the texts systematically. Similarly, what counts as a resolution? Is *smilte katten* "the cat smiled" in example text 1 a resolution? Is this a satisfactory way of ending the story, actually a clever resolution letting the reader draw some conclusions on her own, or is it too vague to qualify for being a resolution? Again, clearly articulated rules that decide the coding are needed. Complex constructs like narrative structure require that the researcher makes several decisions in the operationalization process. When the researcher gives reasons for the choices and is open about

the process, other researchers can judge whether this was an acceptable way of operationalizing the construct. Transparency through strictly formalized operationalizations also makes the coding easier to execute and replicate.

Narrative structure is a feature that demands more interpretation than for example text length and spelling (cf. the previous section). This makes it hard to formalize rules for the coding, and challenging to code objectively. As stories can be formulated in numerous ways, some interpretation is inevitable when a rater scores a specific text according to the coding rules. At the same time, the analysis of narrative structure provides potential for analysing complex aspects of text quality. A measure like spelling gives valuable information, but does not capture all aspects of interest in stories written by children. In a study that claims to assess text quality in beginning writers' stories, a measure of narrative structure would clearly be a relevant measure. The example study discussed in this article sought to find valid measures of text quality – measures that truly reflect quality aspects of the texts, and therefore several measures were included, a holistic quality score and measures of text features on different levels of language.

Since studies show that raters vary in their assessment of text quality, it is recommended to use multiple raters (Bouwer et al., 2015; Huot, 1990a). In this way, the degree of agreement in the coding between the raters can be calculated statistically through the measure of interrater reliability (Bordens & Abbott, 2002). As long as the agreement is acceptable, usually .7 or better, there is reason to believe that the assessments have been done objectively enough. Acceptable agreement indicates that, in spite of some variation, there is consensus about most of the coding. Language makes it possible to make stories of all kinds with different layers of meaning, which makes it unrealistic to achieve 100% agreement in the coding of features like narrative structure. Nevertheless, by accepting some divergence in the coding, it is possible to analyse complex text features within a quantitative frame. Some of the advantages with quantitative research is that it can be used to identify causal and correlational relationships between variables, and also that findings can be generalized when some prerequisites, e.g. related to sampling, are fulfilled (Bordens & Abbott, 2002). If for instance a correlational relationship between modality and text quality in texts by beginning writers is established, it is possible to

predict how students perform when writing in different modalities. Generalized knowledge about factors affecting the writing performance of a specific population, like first graders, is valuable because it can be used in decision-making in the politics of education.

Another example of a construct that could be relevant for the quality of a story, is originality. This was considered for the example study, but not included in the final measures because of the difficulties with operationalization. One reasonable way of capturing this construct could be to compare the storyline of a text to what one could expect as obvious or standard solutions given the task, and operationalize originality as solutions that are not standard, but still relevant. The standard solutions again would have to be defined, e.g. as a composition of fixed narrative phases. This probably would capture instances that clearly would be thought of as original, both by researchers and laymen. However, would this operationalization cover all of the original stories? Also, a concept of originality will typically stretch, or break with, text conventions. This makes it hard to find the boundary between the very creative and the incoherent – for example if one element in the text is so creative that it does not connect to the other elements, or if the whole text is too far from what one would expect given the task and the rest of the context. As human beings we are able to intuitively judge if a story is creative or incoherent. We can also reflect on this decision, and make the inferences and judgments in our interpretation explicit. When making rules for coding, the researcher tries to formalize such interpretations. The operationalization of a construct like originality is difficult, because originality can appear in various forms in texts. Careful descriptions of a construct will increase the chances of an acceptable interrater agreement. However, careful descriptions might also exclude other instances of originality, that are only slightly different. In the process of finding accurate measures there will be a trade-off between different considerations (Cartwright & Runhardt, 2014). When being clear about what is being measured and how, limitations of a study can be illuminated.

In quantitative studies of text quality, measures that do not to any great extent demand interpretation can quite objectively be executed in the sense that different raters will code exactly the same way. However, in some cases analysis of text quality should include features that demand

more interpretation in order to make the analysis more interesting, more complete or more accurate. Interrater reliability can be used to check and document that the agreement of the coding is acceptable, which indicates that there might be a certain amount of interpretation, but not more than what is regarded tolerable. To ensure agreement among raters the language used in the guidelines for the assessment, like holistic rubrics and formalized coding rules used for analytic assessment, has to be precise. Language offers possibilities to describe and explain with complexity and nuance, but language can also be ambiguous. Thus, the researcher is obliged to reflect on language to make clear definitions. Precise guidelines for assessment give potential for assessing aspects of the texts that to a considerable extent demand interpretation. Studies that both apply valid measures of text quality and have acceptable interrater reliability, can provide valuable knowledge about writing performance that can also inform writing instruction practices.

Concluding remarks

This article has discussed the analysis of texts in a quantitative study of writing performance in light of the philosophical traditions of naturalism and interpretivism. A study that investigates texts by beginning writers was chosen as example, as the analysis of text features and quality in these texts is challenging to execute objectively in several ways. It was argued that insights from interpretivism are also valid for this text study based on naturalism. Firstly, on a fundamental level, all text analysis of meaningful structures requires interpretation. The only way to access meaning, e.g. to understand a text, is through the first-person point of view, though one's own sensory apparatus and mental abilities. The researcher is a human being that has to understand the text – make meaning out of it – to be able to quantify the meaningful structures in the text. Further, it was argued that in the process of defining and operationalizing constructs, language plays a crucial role. Rigorous work with language in the guidelines for assessment is decisive for achieving objectivity and transparency in the coding of the texts. In quantitative studies of text quality, the measurement of interrater reliability enables analyses of

features that to a high degree require interpretation, and secures that this is done within the limits of what is considered objective enough. Thus, knowledge about the general characteristics of the text quality in first grader's written stories, can be achieved.

Eivor Finset Spilling
Volda University College
Mailbox 500
NO-6101 Volda, Norway
eivor.finset.spilling@hivolda.no

Literature

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Lawrence Erlbaum Associates.

Berninger, V., Yates, C., Cartwright, A., Rutberg, J., Remy, E., & Abbott, R. (1992). Lower-level developmental skills in beginning writing. *Reading and Writing: An Interdisciplinary Journal*, *4*(3), 257–280. https://doi.org/10.1007/BF01027151

Bordens, K. S., & Abbott, B. B. (2002). *Research design and methods: A process approach* (6th ed.). McGraw-Hill.

Bouwer, R., Béguin, A., Sanders, T., & Van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, *32*(1), 83–100. https://doi.org/10.1177/0265532214542994

Bryman, A. (1984). The debate about qualitative and qualitative research: A question of method or epistemology? *The British Journal of Sociology*, *35*(1), 75–92. https://doi.org/10.2307/590553

Cartwright, N., & Runhardt, R. (2014). Measurement. In N. Cartwright & E. Montuschi (Eds.), *Philosophy of social science: A new introduction* (pp. 265–287). Oxford University Press.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Dockrell, J. E., Connelly, V., Walter, K., & Critten, S. (2015). Assessing children's writing products: The role of curriculum based measures. *British Educational Research Journal*, *41*(4), 575–595. https://doi.org/10.1002/berj.3162

Gadamer, H. G. (2004). *Truth and method*. (J. Weinsheimer & D. G. Marshall, Trans.) (2nd rev. ed.) Continuum. (Original work published 1960)

Gamlem, S. M., Rogne, W. M., Rønneberg, V., & Uppstad, P. H. (2020). Study protocol: DigiHand – the emergence of handwriting skills in digital classrooms. *Nordic Journal of Literacy Research*, *6*(2), 25–41. https://doi.org/10.23865/njlr.v6.2115

Gorton, W. A. (2010). The philosophy of social science. In *Internet Encyclopedia of Philosophy*. http://www.iep.utm.edu/soc-sci/

Grabowski, J., Becker-Mrotzek, M., Knopp, M., Jost, J., & Weinzierl, C. (2014). Comparing and combining different approaches to the assessment of text quality. In D. Knorr, C. Heine, & J. Engberg (Eds.), *Methods in writing process research* (pp. 147–165). Peter Lang.

Haugen, T. A. (2021). Om språkvitenskapen som forståelsesbasert, analytisk vitenskap: Et metateoretisk essay. In T. A. Haugen, S.-A. Myklebost, S. J. Helset, & E. Brunstad (Eds.), *Språk, tekst og medvit* (pp. 11–46). Cappelen Damm Akademisk.

Hayes, J., & Flower, L. (1980). Identifying the organization of writing processes. In L. N. Gregg, & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Lawrence Erlbaum Associates.

Huot, B. (1990a). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, *41*(2), 201–213. https://doi.org/10.2307/358160

Huot, B. (1990b). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, *60*(2), 237–263. https://doi.org/10.3102/00346543060002237

Labov, W., & Waletzky, J. (1967). Narrative analysis: Oral versions of personal experience. In J. Helm (Ed.), *Essays on the verbal and visual arts. Proceedings of the 1966 annual spring meeting of the American ethnological society* (pp. 12–44). University of Washington Press.

MacArthur, C. A., & Graham, S. (2016). Writing research from a cognitive perspective. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed.) (pp. 24–40). Guildford.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Palgrave Macmillan.

Martin, J. R., & Rose, D. (2008). *Genre relations: Mapping culture*. Equinox.

Moran, D. (2002). *Introduction to phenomenology*. Routledge.

Peterson, C., & McCabe, A. (1983). *Developmental psycholinguistics: Three ways of looking at a child's narrative*. Plenum Press.

Roberts, C. W. (2000). A conceptual framework for quantitative text analysis. *Quality and Quantity*, *34*(3), 259–274. https://doi.org/10.1023/A:1004780007748

Spilling, E. F., Rønneberg, V., Rogne, W. M., Roeser, J., & Torrance, M. (2021). Handwriting versus keyboarding: Does writing modality affect quality of narratives written by beginning writers? *Reading and Writing: An Interdisciplinary Journal*. https://doi.org/10.1007/s11145-021-10169-y

Van Steendam, E., Tillema, M., & Rijlaarsdam, G. (2012). Introduction. In E. Van Steendam, M. Tillema, G. Rijlaarsdam, & H. Van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practice* (pp. ix–xxi). Brill.