# From Standards to Instruments: An Ethnographic Study of Four Dimensions of Sociotechnical Translation in NLP Development

**Johanna Fischer**
HafenCity University
johanna.fischer@hcu-hamburg.de

## Abstract

This article offers a sociological perspective on the process of developing NLP systems for text simplification, informed by science and technology studies (STS). Using the example of a specific simplification system in an applied research project for text simplification according to DIN standards, I open the "black box" of development. I propose four dimensions for analyzing the translation of norms into technological systems: data, context, epistemology, and ontology. The analysis reveals that every technical decision is also a sociotechnical translation: a decision regarding who speaks (data), which situations are deemed relevant (context), which knowledge is treated as valid (epistemology), and which reality is produced by the instruments (ontology). For NLP practitioners, these dimensions offer a heuristic for interrogating dataset choice, context transfer, standards operationalization, and metric-based evaluation rather than treating them as purely technical steps.

## 1 Introduction

Research into measuring and simplifying linguistic complexity has a long history. Since the 1920s, researchers have been developing metrics and formulas to define characteristics that represent the complexity of language (Zamanian and Heydari, 2012). The roots of this research lie in pedagogy and psychology. The initial questions focused on teaching language, literacy, and writing, with the aim of improving reading and comprehension skills. Over time, a huge interdisciplinary field has evolved combining pedagogy, linguistic, psychology, media research, and computer science.

Plain Language (German "Einfache Sprache") is a concept that defines a certain level of comprehension. There are different levels defined for German: scientific language, plain language, and easy language. Plain language is a written language intended for native speakers. Some equate it with level B1, but that is an EU proficiency framework designed for learners (Baumert, 2023). Plain language is used to reduce barriers to comprehension, effective communication, inclusion, and the transfer of knowledge to laypeople. In Germany, relevant requirements have been codified since 2024 in the DIN[1] documents DIN ISO 24495-1 and DIN 8581-1 (DIN e. V., 2024).

The DIN documents specify how "plain language" must be structured to be considered understandable by the target audience. The rules serve as a guide for authors to tailor their texts to specific needs. Thus, DIN standards for plain language are written for people who interpret them for their target group. The standards are not operationalized for computers. Therefore, to make them function as deterministically as computers require, other studies have attempted to define clear metrics that can be used to determine whether a text complies with the rules or not (Decher and Dembach, 2025). Caution is warranted here, as the impact of certain rules on subjective perception of complexity, such as sentence length (Wolfer et al., 2015) or sentence topology (Wöllstein-Leisten, 2014), is a subject of controversy in the academic literature. Disciplines within NLP are growing increasingly aware of the context sensitivity of their research, as demonstrated by the KONVENS 2026 conference theme ‘Context Matters’.

Related work in NLP has already responded to concerns about transparency and responsibility by proposing formats for Data and Model documentation (Holland et al., 2018; Bender and Friedman, 2018; Gebru et al., 2021; Mitchell et al., 2019). These initiatives show that questions of provenance, intended use, limitations, and accountability are already being negotiated within the field. Rather than proposing another documentation format, this

[1] Deutsches Institut für Normung: German Institute for Standardization

paper contributes a reflexive heuristic for making visible how and which social norms are translated into concrete development decisions across inputs, situational embeddings, knowledge, and effects.

Large Language Models (LLMs) nowadays promise a scalable, automatic translation to plain language (Maaß et al., 2025). The challenge has shifted from human resources and cognitive abilities to technical implementation issues. One example is the question of how to get an LLM to reliably convert text into plain language (Anschütz et al., 2025). There are various ways to achieve this, including rule-based systems, prompt engineering, and model chains, that is, approaches that encode constraints either as explicit rules, as prompt instructions for an LLM, or across multiple processing steps (e.g. Janfada and Minaei-Bidgoli (2020); Wei et al. (2020)).

Classical instruments for complexity assessment such as the Flesch-Kincaid Grade Level uses only three features to assess the readability of a text corpus (word count, sentence count, and syllable count) (Kincaid et al., 1975).

$$\text{Grade Level} = 0.39 \times \left(\frac{\text{Total Words}}{\text{Total Sentences}}\right) + 11.8 \times \left(\frac{\text{Total Syllables}}{\text{Total Words}}\right) - 15.59$$

Newer studies suggest that one broadens the repertoire of features (Seiffe et al., 2022; Mohtaj et al., 2022). One of these studies used up to 255 features (Lee et al., 2021). Others build suites of instruments to assess text with varying target measures (Graesser et al., 2011; Crossley, 2025). However, there is no consensus on which metric is relevant for which target group (Deane et al., 2006, p. 257). Although DIN standards prescribe very specific criteria, datasets annotated with subjective perceptions of complexity show that people define what is complex to them in very different ways (Mohtaj et al., 2022; Seiffe et al., 2022).

In the applied sciences, this normative uncertainty is typically circumvented pragmatically by using established datasets as *ground truth* [2], common metrics are repurposed, and the alignment of their results with one's own values is taken as proof of the system's validity. What often recedes from view in this process is a chain of normative decisions that lead to a deterministic system of classification and standardization (Bowker and Star, 2000).

Against this backdrop, the aim of this article is to offer a sociological perspective on the development process, informed by concepts from science and technology studies (STS). This approach allows us to observe critical moments of sociotechnical translation of standards into algorithms within the work process. Translation is used here in the sense of Actor-Network Theory (ANT), that is, as the practice of creating a link between two entities that were not linked before. If this translation is successful, the main actor (in this case a technological instrument) can speak for others in its own tongue and thus receives "human properties, and is enrolled in and mobilized within the collective" (Shiga, 2007, 42). Using the example of a specific AI system in an applied research project, I open the "black box" of development. I call it a black box because, typically, once a product is completed, traces of uncertainty and conflict are concealed beneath a veil of linear explanations and the rational naturalization of technical decisions. However, my research reveals development in the midst of the action, its indeterminate uncertainty, and its contingent possibilities.

I follow a twofold objective: First, I aim to describe normative decisions made in a specific applied research project on the automation of natural language. Using the ethnographic approach of participatory observation, I highlight key milestones that have shaped the system. Second, I propose conceptual dimensions for questioning the process. This constitutes a methodological contribution to reflexive research during the development of software systems. For NLP practitioners, this contribution is not a finished implementation protocol, but a way of locating where normative choices enter routine development work: in dataset selection, contextual assumptions, standards operationalization, and evaluation design. To this end, the questions addressed in this article are divided as follows:

- **RQ 1:** What normative decisions can be observed?
- **RQ 2:** How can these normative decisions be systematically identified and analyzed?

To answer these questions, I differentiate four

[2] Jaton explains the term for his STS audience as follows: "referential repositories that work as material bases for algorithms. The centrality of ground truths and of the work required to build them make me assert that, to a certain extent, we get the algorithms of our ground truths" (Jaton, 2020, 24)

dimensions for the translation of norms into technological systems: data, context, epistemology, and ontology. The claim is not that one case exhausts all possible development settings, but that these dimensions provide a portable analytic heuristic for similar NLP projects.

On the one hand, the article raises awareness of the normative implications of seemingly rational technological decisions. I demonstrate that decisions regarding data entrench certain assumptions about prospective users, but that the context of data generation also "creeps in." By examining the development of specific metrics, I highlight the reductionist approach to language and how the system constrains complexity in a standardized manner. By examining the evaluation of system outputs, I highlight the agential cut made by the researcher, as well as the construction of a convincing narrative that naturalizes decisions as technical necessities. On the other hand, I offer conceptual perspectives to make visible the relationships between ground truth, metrics, and AI systems beyond established correlation-based validation. For NLP practitioners, the four dimensions can be translated into guiding questions for reflecting on dataset choice, context transfer, standards operationalization, and metric interpretation. This helps document why a given design decision was made and whose assumptions it stabilizes.

### 1.1 Positioning

This analysis is the result of participatory observation. I am a trained sociologist of technology and, during the observation period, was working on the project as a data scientist following a career change. In the project, I was responsible for the workflow described here. As a newcomer, I relied on the advice and methods of the project leader. From an STS perspective, this position is not a disadvantage. It afforded me a specific vantage point between two worlds: that of the outsider and that of the insider. I documented my observations in a field diary, which I draw upon in describing this case. The four dimensions were not derived from this diary alone in a purely inductive manner. Rather, the analysis was informed by a broader line of STS research on the development of digital tools, including my published work on the making of technological instruments in machine learning (Fischer, 2025). Across this and the present case, I traced recurring interactions around data work, technical problems and repairs, shared standards and assumptions, and imagined scenarios of use. These observations were organized into recurring clusters according to the methods of Grounded Theory (Glaser and Strauss, 2017) in several iterations concerning data practices, contextual scenarios of use, knowledge orders, and the material effects of instruments. In a final iteration, these clusters were further sharpened through STS concepts attentive to epistemic and ontological questions. Comparative insights from my ongoing dissertation research further informed this analytical consolidation. The four dimensions presented here – data, context, epistemology, and ontology – are offered as an analytic heuristic rather than as an exhaustive typology.

At the same time, my position raises questions of accountability. This text is not the result of detached scientific observation, but largely the result of self-observation. It is both a documentation and a critique of my own practices. Furthermore, the grouping of my observations into four dimensions is just as much a normative operationalization as the implementation of DIN standards in instruments. It must be said that methods are not objective tools for making the world "out there" visible. On the contrary, methods shape the world, and it is therefore appropriate that we reflect seriously on them (Beck et al., 2012, 16). Therefore, the dimensions serve as a model-based aid for reflexive research and are intended to help achieve a shift in perspective in order to view the technological object as it is emerging (Jensen, 2020, 110–111).

## 2 Case: Rule-aware Simplification

This section provides a brief technical overview of the system. The applied research project[3] involved a workflow for translating complex texts into plain language. The system fulfills two core functions: first, the automatic identification of sentences that need to be simplified by predicting users' subjective perception of complexity; second, the generation of simplified sentences using an LLM that considers specific DIN requirements for plain language, which I called the Rule-aware Model. I report technical details here only insofar as they are analytically relevant for the sociotechnical argument. For the text complexity prediction function, I used a pre-built tool which I will refer to as the Text-Complexity Regression.

[3]The details of the project must remain anonymous for privacy reasons.

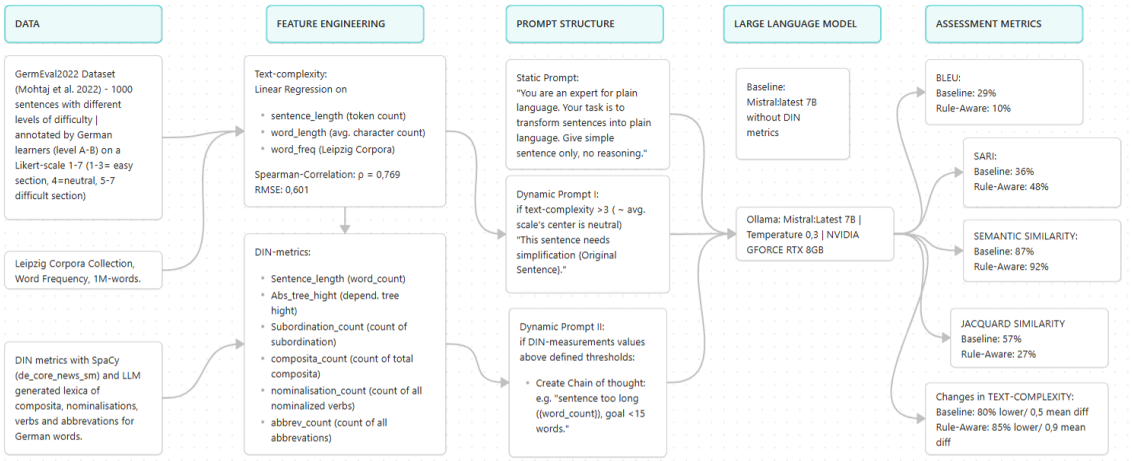


Figure 1: Flowchart of the system architecture, from data and feature engineering to prompt structure, the LLM, and assessment.

### 2.1 First Functionality – Complexity Prediction

As a reference for subjective complexity assessment in text, I was forwarded a training and test dataset from a public challenge called GermEval 2022 (Mohtaj et al., 2022). The dataset is an adapted excerpt from a more detailed set called TextComplexityDE19 (Naderi et al., 2019).

For the first functionality, I used a pre-built tool from a previous project my teamleader had worked on: *Text-Complexity Regression*. The tool uses a feature set which is common for text assessment: basic text features (*word_length*, *sentence_length*) as a proxy for established readability formulas, and word frequency (*uniqueness*) as a proxy for word familiarity. The basis for measuring word familiarity comes from the Leipzig Corpora Collection (a German Wikipedia corpus containing 1 million sentences) (Leipzig, 2021). These three features – word_length, sentence_length, and uniqueness – serve as data points for training a linear regression model in the training set, which predicts mean ratings (Likert scale 1–7). When applied to the test set, the Text-Complexity Regression achieves an RMSE[4] of 0.601 and a Spearman correlation[5] of 0.769. Compared to other results in the GermEval 2022 challenge, the values seem solid considering the simplicity of the tool.

[4]Root Mean Square Error (RMSE) measures the average prediction error, penalizing larger deviations more strongly; lower values indicate closer predictions.

[5]Spearman correlation measures how well the model preserves the ranking of examples rather than their exact values; higher values indicate that the relative order of complexity judgments is captured more accurately.

For sentences classified by the Text-Complexity Regression as requiring simplification (predicted value > 3.0), a prompt for a large language model was rule-based generated. Although the original annotation scale ranges from 1 to 7, the mean opinion scores in the dataset cluster effectively between 1 and 6, with 7 scarcely occurring after averaging across raters. In the project team, the interval between 3 and 4 was therefore treated as the neutral center of the score distribution, so that a threshold above 3.0 was used as a conservative decision threshold.

### 2.2 Second Functionality – DIN-aware Prompting

My colleagues and I set out to incorporate summaries of two German standards for plain language, DIN 8581-1 and DIN ISO 24495-1 (DIN e. V., 2024). Six requirements were extracted for plain language and converted into measurable indicators inspired by earlier approaches to compute the DIN norms for plain language (Decher and Dembach, 2025). These were transformed into prompt templates. Each prompt consists of three components:

1. System instruction (static).
2. Original sentence (dynamic, only if the criteria are violated: text complexity > 3).
3. DIN adjustment (dynamic, if DIN thresholds are violated).

The DIN violations are converted into templates and included in the prompt as a Chain of Thought (CoT): e.g. "Sentence too long with 18 words,

target < 15." The system used a locally hosted Mistral:latest 7B model. Its temperature hyperparameter was set to 0.3 to encourage relatively deterministic generation.

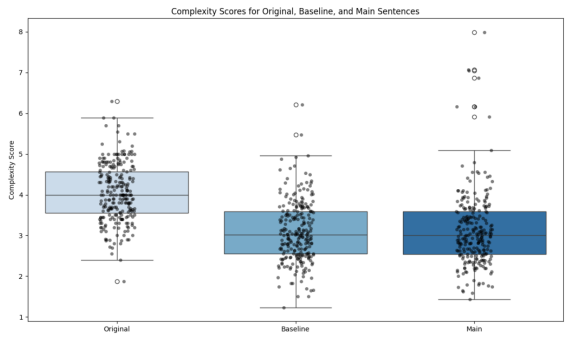


Figure 2: Distribution of text complexity in Original data, Baseline and Rule-aware Model (Main).

### 2.3 Assessment of the system's quality

An excerpt from the TextComplexityDE19 dataset served as the test set. In addition to the original sentences and complexity scores, the test set includes 250 manually simplified sentences for comparison with the output of the LLM.

Table 1 reports the assessment metrics for the Baseline and the Rule-aware Model. As expected, BLEU[6] and comparable metrics (e.g., Jaccard similarity[7]) are low, indicating significant variation in the words between the original and simplified versions. However, the semantic similarity calculated using the BERT-similarity[8] Transformer model suggests a high degree of content similarity. SARI[9] achieves a moderate value, indicating significant overlaps between the Rule-aware Model's results and the reference sentences. The reductions in predicted text complexity, by contrast, refer to the difference from the original sentences and are visualized in Figure 2 across all three conditions (Original, Baseline, and Main = Rule-aware Model). This comparison shows an average reduction of 0.9 scale points for the baseline and 1.3 points for the Rule-aware Model. It also illustrates that CoT, when combined with DIN guidelines, reduced the predicted text complexity of the simplification.

| Metric | Baseline | Rule-aware Model |
|---|---|---|
| BLEU | 0.287 | 0.1095 |
| Jaccard-Similarity | 0.571 | 0.2684 |
| BERT-Similarity | 0.870 | 0.9243 |
| SARI | 0.3581 | 0.4820 |
| Text-Complexity | 3.1 | 2.7 |

Table 1: Assessment Metrics

## 3 Analysis: Four Dimensions

At first glance, this linear representation of the system seems convincing. But let us now turn our attention to the moments where translation shapes the system's configuration. In this section, we will examine the configuration across four dimensions: data, context, epistemology, and ontology.

### 3.1 Ground truth: Translating subjective human evaluations

Data are not neutral raw materials; rather, they bear the traces of the conditions under which they were collected, the logics of their classification, and the institutional contexts in which they were produced (Thylstrup, 2022). There are no untouched raw forms of data (Gitelman, 2013). Others emphasize: "Data [...] often figure as an element of already ongoing, much older practices of classifying, categorizing, quantifying, and producing actionability on the social" (Velkova and Johnson, 2025, 7). Despite this, in applied projects, data is often considered the 'raw' input, representing ground truths from an outside world. Thylstrup points out that, despite their central role in algorithmic technology, datasets rarely become the subject of investigation themselves (Thylstrup, 2022, 655). Based on the arguments above, the choice of a ground truth is not a purely technical preliminary decision, but an ethical-political one: in the use case, it determines who is considered the defining authority of "complexity". At best, carefully considered selection of data accurately reflects many people while also taking their limitations into account; at worst, it leads to the exclusion of entire groups in a society and ignores bias and normalization (D'Ignazio and Klein, 2020). Therefore, a critical study must focus on the characteristics of the data, such as "data forms, processes, and purposes, and consider what is at stake when data are produced and deployed to

[6]BLEU measures n-gram overlap between source and generated text; high scores therefore privilege lexical similarity rather than comprehensibility as such (Papineni et al., 2002).

[7]Jaccard similarity captures the proportion of shared words between original and generated text and thus treats lexical overlap as a proxy for similarity.

[8]BERT-similarity estimates semantic similarity between texts on the basis of contextualized token representations, making content preservation legible even when wording changes substantially (Zhang et al., 2020).

[9]SARI evaluates simplification by comparing which n-grams are added, kept, or deleted relative to human reference simplifications (Xu et al., 2016).

create knowledge, manage society, derive business value, and perform numerous other tasks" (Kitchin, 2024, 2).

Related work in NLP has already responded to comparable concerns by proposing documentation practices such as Data Nutrition Scores, Data Statements and Datasheets for Datasets, which are intended to make datasets more transparent with regard to provenance, intended use, and limitations (Holland et al., 2018; Bender and Friedman, 2018; Gebru et al., 2021). My case does not challenge the value of such documentation; rather, it shows its practical limit. In the project studied here, relevant information about the dataset was in fact available across repositories, accompanying papers, and metadata, yet it did not meaningfully enter development decisions until it became the object of this reflexive analysis. The problem, then, is not only missing documentation, but the absence of practices that explicitly justify and document why particular datasets are selected in the first place.

The data selected as ground truth – particularly the annotations regarding complexity – serve as an inscription[10]. This provides the fundamental record of traces of subjective assessment, which will be translated into its first functionality: Text-Complexity Regression. In the present system, the GermEval 2022 dataset was chosen as ground truth. This dataset contains complexity ratings from German learners (levels A–B) regarding Wikipedia sentences. The decision to use this dataset was based on pragmatic considerations: Since it had already been used in a similar project—from which the text complexity regression was also adopted—this instilled confidence that both the data and the algorithm had already been sufficiently evaluated.

From an STS perspective, what is now crucial is what this choice obscures: the plurality of potential target audiences for plain language. The complexity ratings given by annotators and future users of the project may differ significantly. The complexity metric does not learn "complexity." It learns to predict the rating of mean complexity for a very specific group under specific circumstances: German learners, in interviews, with regard to Wikipedia texts. In this form, the value of the mean opinion score variable *mos_complexity* was very quickly regarded as an established fact by the team. The relationship between annotators, learner level, reasons for perceived complexity, and data collection disappeared behind the clean mean value. Additionally, the factor of legitimation through previous projects comes into play here. Choosing a dataset that had been used before meant that established conditions were replicated. I used the dataset without question and didn't challenge the recommendation until I encountered issues later in the project when I used the complexity metric on real transcripts.

The weight assigned to the word frequency feature in the Text-Complexity Regression translates another characteristic: text genre. The model has learned that words with high frequency in the respective genre (Wikipedia article) are an indicator of "simplicity." It can thus be a useful tool in the context of Wikipedia articles. The heuristic of the pre-built tool penalizes every word not included in the dictionary by assigning it the lowest frequency value. Word frequencies can vary greatly in spoken language and everyday conversation. When applied to transcripts of political debates in the Bundestag, I found that formal greetings such as "Geehrte Abgeordnete!" ("Honorable members of parliament!") are marked as very complex structures.

It turns out that the heuristics underlying Text-Complexity Regression translate a highly specific perspective: the tool assumes a group of people (German levels A–B) and collected scores for the purpose of assessing learning materials. Additionally, the metric adopts a highly reductionist approach with only three values, translating only these as relevant features for text assessment. Word frequency poses another problem. Not only are words from Wikipedia articles rewarded, but unfamiliar words are "penalized" with the highest value, which renders the metric unusable for formal spoken language, such as in the transcribed debate from the federal parliament.

### 3.2 Context: Open scenarios as a contradiction to design justice

In the present system, the issue of contextual fit concerns two levels: the origin of the pre-built instruments used and their intended location of use in the new project. Sasha Costanza-Chock's work on design justice (Costanza-Chock, 2020) highlights that technological development rarely takes into account the actual contexts of use of those for whom it is intended. Inclusion is often treated as an afterthought, rather than as a constitutive dimension

[10] Inscription refers to the process of recording traces of an object or subject. These traces get translated into diagrams, datapoints, and numbers that are intended to reflect material's specific properties. The term was coined by Latour and Woolgar in their book *Laboratory Life* (Latour and Woolgar, 1986, 45).

of the development process. Costanza-Chock therefore considers "Design [...] [as] a way of thinking, learning, and engaging with the world" (Costanza-Chock, 2020, 15). In the practical implementation of design decisions, abduction – that is, deriving the best prediction from incomplete observations – stands in contrast to the scientific concepts of deduction and induction (Aliseda, 2005). The design of specific products will therefore always be largely speculative. It is evident that the imagination that shapes such technological speculations has a strong influence on their form (Wucherer, 2025, 112). This goes hand in hand with an ethical responsibility and the awareness that design is always a form of "world-making" (Escobar, 2018). The perspectives that, through abduction, outline the imagined context for a future tool are therefore largely decisive in determining what kind of world the finished object will bring into being.

From the perspective of Design Justice, the absence of prospective users of this tool is problematic: In this case, there was not even a clear definition of a target group. The only context given was 'loose sketches of possible scenarios': police talking to people from heterogeneous backgrounds; school consulting parents; entrepreneurs managing documents and files. So what to translate in such a case?

I simply imported context from other research: The system's feature set combines features from three sources: from established readability formulas (Flesch-Kincaid), word frequency (Wikipedia corpus), and operationalized DIN criteria. Each of these instruments was developed in a specific context and validated for specific purposes. The readability formula, for example, originated in U.S. military research in the 1970s (Kincaid et al., 1975) and was developed for English-language technical manuals. The score measures the time a soldier needs to read texts containing familiar technical vocabulary (Graesser et al., 2011, 224). Its application to German Wikipedia sentences evaluated by German learners constitutes a transfer across multiple contexts and cultures.

For other parts, I reduced context to a minimum and considered complexity to be computable. The operationalization of the sentence-length rule (binary indicator: 1 if >15 words) is an example of a lack of context sensitivity. A sentence with 14 words is considered less problematic, even if it contains five compound words and three nominalizations; a sentence with 16 words but otherwise simple structure is considered in need of simplification. This logic is embedded in the system; it is reproduced in every application, regardless of the specific context. A context-sensitive system would need to allow for dynamic thresholds or weighted combinations of features.

### 3.3 Epistemology: Standards and Classification as Systems of Knowledge

The work of Star and Bowker (Bowker and Star, 2011, 2000) has shown that classification systems and standards are not neutral technical tools, but rather *systems of knowledge* that privilege certain perspectives and render others invisible.

"Classifications should be recognized as the significant site of political and ethical work that they are. They should, in a word, be reclassified as key sites of work, power, and technology." (Bowker and Star, 2011, 147)

The DIN standards for plain language are one such system of knowledge: they define what constitutes "plain language." However, their format (continuous text) is not particularly prescriptive. The DIN serve as a boundary object – an object that provides a structuring framework across different disciplines, contexts, and activities. "Boundary objects are objects which are both plastic enough to adapt to local needs and constraints of the several parties employing them, yet robust enough to maintain a common identity across sites. They are weakly structured in common use, and become strongly structured in individual-site use." (Star, 1992, 406) In our system, this interpretable but ambiguous standard is operationalized and thus translated into a technically unambiguous form (Pütz, 2021). This translation into measuring instruments has significant epistemological implications, as it has a determinative effect: it simplifies heuristics, flattens meaning, and erases contextual traces.

The DIN standard is flexible enough to be interpreted differently by various stakeholders (linguists, users, developers) and robust enough to serve as a common reference across these different contexts. In the present system, however, the standard is not left as a boundary object but is closed off: it is translated into binary and numerical indicators whose interpretation is unambiguous (e.g. sentence length >15 words = DIN violation). This closure eliminates the interpretive flexibility inherent in the standard as a boundary object. The operationalization of the compound word rule highlights the problem with closed definitions: The DIN stan-

dard recommends avoiding compound words or writing them with a hyphen. In the system, this is translated as a binary indicator: compound words without a hyphen are considered a violation. This operationalization excludes several possible interpretations: (a) the possibility that some compound words (such as established ones like "Hundehütte" [doghouse]) are acceptable, (b) the possibility that the length of the compound plays a role, (c) the possibility that the context (such as the accumulation of several compound words) is decisive. The system does not recognize these distinctions.

Another example of standards is the correlation with established readability metrics from the field. In the documentation for Text-Complexity Regression, a correlation of the tool is interpreted as validation: the Text-Complexity Regression actually measures complexity because it correlates with the Flesch-Kincaid Grade Level. From an STS perspective, this is an epistemic fallacy: correlation is being confused with validity. The Flesch-Kincaid metric was developed for English-language technical military manuals and operationalizes text complexity through sentence count and syllable count. The fact that results from Text-Complexity Regression correlates with it does not show that it measures complexity correctly, but rather that it performs similar reductions to a metric developed for an entirely different context.

### 3.4 Ontology: World-Making Instruments

Measuring instruments are not merely representations of an independent reality, but *agential cuts* that performatively bring reality into being. Instruments divide the world into measurable and non-measurable phenomena; in doing so, they constitute the very phenomena they claim to measure (Barad, 2007). In this system, the technical tools (the feature set, the Text-Complexity Regression, the assessment metrics) are not neutral observational instruments, but rather active shapers of what manifests as "complexity." As statistical metrics, they translate selected characteristics of phenomena into numerical values. In doing so, these measured values take on a dual character - they generate factuality for phenomena that elude human perception and simultaneously present themselves as perceivable numerical values (Heintz, 2012, 7).

The whole feature set comprising readability features and metrics derived for the DIN standards can be regarded as agential cuts. This feature set translates what is measured as properties of complexity in the text: sentence length, word frequency, syntactic depth, number of compound words, nominalizations, and abbreviations. Anything not represented by these features (e.g., cohesion or different text genres) becomes invisible due to these agential cuts. Many phenomena are thus excluded once this feature set is established as a measuring instrument.

The BLEU and BERT metrics (Table 1) represent another agential cut. Furthermore, through their performance alongside developers, BLEU and BERT function as legitimizing agents. BLEU measures n-gram overlap with a reference; it "knows" nothing about comprehensibility, readability, or suitability for the target audience. When the system is evaluated using BLEU, similarity becomes the benchmark. BLEU embodies a specific ontology of quality (high similarity to the original). In the developers' analysis, this is further altered by the original expectation regarding BLEU scores: because the BLEU score is low (little n-gram overlap), the LLM must have altered the sentence significantly.

Quality comes down to how BLEU is interpreted: little overlap is interpreted as greater simplification. When combined with the other scores, the metrics reveal their persuasive power. This is because BERT-similarity measures content similarity. Since the BERT-similarity is high, the developers interpret it as a sign that although the BLEU score is low (meaning that words have been significantly altered), the content has nevertheless been preserved.

## 4 Conclusion

The four dimensions demonstrated here enable a structured reconstruction and reflection of technological development. They reveal that every technical decision is also a sociotechnical translation: a decision regarding who speaks (data), which situations are deemed relevant (context), which knowledge is treated as valid (epistemology), and which reality is produced by the instruments (ontology). Documenting these dimensions is the first step toward a reflexive practice that is aware of its own conditions. The dimensions therefore do not constitute a definitive critique, but rather serve as a tool for reflexive research practice. They allow normative decision-making moments to be addressed systematically during project implementation.

At the same time, the contribution of this paper should not be overstated. A single STS case study cannot provide a universal implementation protocol for integrating these dimensions into practice.

Translating them into concrete workflows, documentation templates, or evaluation routines is a subsequent task for practitioners working in specific application settings. What this paper can provide is a clearer vocabulary for identifying where such practical work needs to begin: at dataset choice, context assumptions, standards operationalization, and in the selection and interpretation of measuring instruments. In practical terms, this means asking whose judgments are sedimented in the data, which use contexts are being imported or excluded, how standards are converted into measurable indicators, and what form of quality a chosen metric makes visible. This is also where the present argument may complement existing documentation initiatives in NLP. If formats such as Model Cards are intended to make the capabilities and limitations of machine learning systems more transparent (Mitchell et al., 2019), then the four dimensions proposed here can support the reflexive work required to produce such documentation. They help to look beyond the formally planned project frame and to articulate the decisions, assumptions, and branching points that emerge during development but are often naturalized afterward. Their application will not result in a perfect technique free from normative assumptions. However, it can contribute to a more reflexive practice that makes these assumptions more transparent and open to change, thereby resisting the temptation to naturalize technical decisions as mere practical constraints.

## References


Atocha Aliseda. 2005. *Abductive Reasoning: Logical Investigations Into Discovery and Explanation*. Springer, Dordrecht and London.

Miriam Anschütz, Silvana Deilen, and Ekaterina Lapshinova-Koltunski. 2025. Künstliche Intelligenz: Wie funktioniert KI für Einfache Sprache? In Christiane Maaß, Ekaterina Lapshinova-Koltunski, and Julian Hörner, editors, *Einfache Sprache mit KI-Tools: Ein Leitfaden für die redaktionelle Praxis*, pages 57–71. Springer Fachmedien, Wiesbaden.

Karen Barad. 2007. *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Duke University Press, Durham, NC.

Andreas Baumert. 2023. *Einfache Sprache: Verständliche Texte schreiben*. Spaß am Lesen.

Stefan Beck, Jörg Niewöhner, and Estrid Sørensen. 2012. *Science and technology studies: eine sozialanthropologische Einführung*. Band 17. Transcript, Bielefeld.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Geoffrey C. Bowker and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences*. The MIT Press, Cambridge, MA.

Geoffrey C. Bowker and Susan Leigh Star. 2011. Invisible Mediators of Action: Classification and the Ubiquity of Standards. *Mind, Culture, and Activity*, 7(1-2):147–163.

Sasha Costanza-Chock. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need*. The MIT Press, Cambridge, MA.

Scott A. Crossley. 2025. Developing Linguistic Constructs of Text Readability Using Natural Language Processing. *Scientific Studies of Reading*, 29(2):138–160.

Paul Deane, Kathleen M. Sheehan, John Sabatini, Yoko Futagi, and Irene Kostin. 2006. Differences in Text Structure and Its Implications for Assessment of Struggling Readers. *Scientific Studies of Reading*, 10(3):257–275.

Sophie Decher and Michael Dembach. 2025. Automatic Generation of Comprehensible German Summaries for Domain-Specific Scientific Definitional Texts. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 202–211, Hannover, Germany. HsH Applied Academics.

Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism*. The MIT Press, Cambridge, MA.

DIN e. V. 2024. *Einfache Sprache: DIN 8581-1 und DIN ISO 24495-1*. DIN Media, Berlin.

Arturo Escobar. 2018. *Designs for the Pluriverse: Radical Interdependence, Autonomy, and the Making of Worlds*. Duke University Press, Durham, NC.

Johanna Fischer. 2025. Hands for AI. Programming in Machine Learning as labor, work, and action. *Grounding Digitalization. Technologies, Materialities, and Spaces*, pages 237–254.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Lisa Gitelman, editor. 2013. *"Raw Data" Is an Oxymoron*. MIT Press, Cambridge, MA.

Barney G. Glaser and Anselm L. Strauss. 2017. *Discovery of Grounded Theory: Strategies for Qualitative Research*. Routledge.

Arthur C. Graesser, Danielle S. McNamara, and Jonna M. Kulikowich. 2011. Coh-Metrix : Providing Multilevel Analyses of Text Characteristics. *Educational Researcher*, 40(5):223–234.

Bettina Heintz. 2012. Welterzeugung durch Zahlen Modelle politischer Differenzierung in internationalen Statistiken, 1948-2010. *Soziale Systeme*, 18(1 + 2):7–39.

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards.

Behrooz Janfada and Behrouz Minaei-Bidgoli. 2020. A Review of the Most Important Studies on Automated Text Simplification Evaluation Metrics. In *2020 6th International Conference on Web Research (ICWR)*, pages 271–278.

Florian Jaton. 2020. *The constitution of algorithms: ground-truthing, programming, formulating*. The MIT Press, Cambridge, MA.

Torben Elgaard Jensen. 2020. Exploring the Trading Zones of Digital STS. *STS Encounters*, 11(1):89–116.

J. P. Kincaid, Jr Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Report No. RBR-8-75.

Rob Kitchin. 2024. *Critical Data Studies: An A to Z Guide to Concepts and Methods*. Polity, Hoboken, NJ.

Bruno Latour and Steve Woolgar. 1986. *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wortschatz Leipzig. 2021. deu_wikipedia_2021.

Christiane Maaß, Ekaterina Lapshinova-Koltunski, and Julian Hörner, editors. 2025. *Einfache Sprache mit KI-Tools: Ein Leitfaden für die redaktionelle Praxis*. Springer Fachmedien, Wiesbaden.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, New York, NY, USA. Association for Computing Machinery.

Salar Mohtaj, Babak Naderi, and Sebastian Möller. 2022. Overview of the GermEval 2022 Shared Task on Text Complexity Assessment of German Text. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, pages 1–9, Potsdam, Germany. Association for Computational Linguistics.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019. Subjective Assessment of Text Complexity: A Dataset for German Language.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ole Pütz. 2021. Managing exactness and vagueness in computer science work: Programming and self-repair in meetings. *Social Studies of Science*, 51(6):938–961.

Laura Seiffe, Fares Kallel, Sebastian Möller, Babak Naderi, and Roland Roller. 2022. Subjective Text Complexity Assessment for German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 707–714, Marseille, France. European Language Resources Association.

John Shiga. 2007. Translations: Artifacts from an Actor-Network Perspective. *Artifact*, 1(1):40–55.

Susan Leigh Star. 1992. The trojan door: Organizations, work, and the "open black box". *Systems practice*, 5(4):395–410.

Nanna Bonde Thylstrup. 2022. The ethics and politics of data sets in the age of machine learning: deleting traces and encountering remains. *Media, Culture & Society*, 44(4):655–671.

Julia Velkova and Ericka Johnson. 2025. Critical Data Studies Meet Sociology. *Sociologisk Forskning*, 62(1-2):7–18.

Wei Wei, Bei Zhou, and Georgios Leontidis. 2020. A Hybrid Natural Language Generation System Integrating Rules and Deep Learning Algorithms.

Sascha Wolfer, Sandra Hansen-Morath, and Lars Konieczny. 2015. Are shorter sentences always simpler? discourse level processing consequences of reformulating jurisdictional texts. In Karin Maksymski, Silke Gutermuth, and Silvia Hansen-Schirra, editors, *Translation and Comprehensibility*, volume 72 of *TRANSÜD. Arbeiten zur Theorie und Praxis des Übersetzens und Dolmetschens*, pages 263–287. Frank & Timme, Berlin.

Sebastian Wucherer. 2025. "A Message from our CEO": Sundar Pichai's vanguard visions of global AI futures powered by Google. In Regula Valérie Burri, Hanna Göbel, and Inga Reimers, editors, *Digitale*

*Gesellschaft*, volume 83, pages 111–130. transcript Verlag, Bielefeld, Germany.

Angelika Wöllstein-Leisten. 2014. *Topologisches Satzmodell*. Number 8 in Kurze Einführungen in die germanistische Linguistik - KEGLI. Universitätsverlag Winter, Heidelberg.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Mostafa Zamanian and Pooneh Heydari. 2012. Readability of Texts: State of the Art. *Theory and Practice in Language Studies*, 2(1):43–53.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT.