

Digital Business University of Applied Sciences

Data Science & Business Analytics

FDA93 SP III Module 4 Research project: Data Engineering &

Pre-Processing

Prof. Claudia Baldermann

**Reflection quality in multilingual student writing:
A Rule-Based Approach to Text Analysis**

Study Project

Submitted by Sabine Wildemann

Student ID 190297

Submission date 07.06.2025

Abstract

This study investigates how the quality of student reflections can be assessed through interpretable, rule-based textual features without relying on generative AI models. Drawing on theories of reflective writing and discourse analysis, the study implements a transparent, language-sensitive methodology combining multilingual preprocessing, rule-based feature engineering, and statistical analysis. From a dataset of 264 anonymized reflections collected via the Rflect platform, linguistic and revision-based indicators are extracted. Two hypotheses are tested: (H1) that linguistic features are associated with reflection depth, and (H2) that revisions between initial and final drafts are positively linked to reflective quality. Non-parametric analyses (Kruskal–Wallis, Spearman correlation) confirm H1, showing that deeper reflections exhibit higher word counts, lexical diversity, reasoning vocabulary, and self-referential statements. H2 receives partial support: conceptual expansions, especially increases in reasoning-related terms, correlate more strongly with final quality than structural changes alone. The findings demonstrate that rule-based, language-aware features can serve as transparent proxies for reflection depth, supporting explainable learning analytics and fair multilingual feedback in education.

List of Abbreviations & Figures

Acronym	Definition
AI	Artificial Intelligence
IDG	Inner Development Goals
LA	Learning Analytics
ML	Machine Learning

Table of contents

1. Introduction
2. Literature Review / Theoretical Framework
3. Methodology / How hypotheses are tested
 - 3.1 Data acquisition
 - 3.2 Data integration
 - 3.3 Data cleansing
 - 3.4 Feature Engineering
 - 3.5 Automation & Preprocessing Output
4. Results
 - 4.1 H1: Relationship Between Linguistic Features and Reflection Depth
 - 4.2 H2: Correlation Between Revisions and Reflective Quality
5. Discussion and Conclusions
6. References
7. Appendix

1. Introduction

Reflection is a central component in modern learning processes, particularly in personal and competency development. Educational technologies increasingly integrate reflective writing exercises to help students develop skills such as critical thinking, self-awareness, and collaboration. Assessing the quality of these reflections, however, remains a challenge – especially in a scalable and explainable way.

The edtech startup Rflect supports student development through structured reflection activities based on the Inner Development Goals (IDGs) framework (Inner Development Goals Initiative, 2021). These reflections are written directly on the platform and cover a wide range of competencies. To support feedback and learning progress, there is a growing need to analyze this content and derive insights — without relying on human raters or generative AI models.

Problem Statement

Although recent research has demonstrated the feasibility of automated approaches for assessing reflection quality using explainable machine learning (ML) and linguistic features (Alrashidi, Almujaal, Kadhum, Ullmann, & Joy, 2023), such methods are not yet widely adopted in digital learning tools. Many platforms either still rely on manual reading or on black-box generative AI tools, which are neither fully transparent nor sustainable in educational contexts – especially when multilingual or competency-based frameworks are required.

Rflect provides a unique dataset of student-written reflections, including versions before and after AI-based feedback. However, the platform currently lacks a method to assess the development of reflections using transparent, structured methods. The challenge is to identify whether textual features and content changes can be used to approximate reflection quality in a reproducible way.

Research Question and Hypotheses

Research Question:

How can students' reflection quality be automatically assessed using textual features and interaction data – without relying on generative AI?

Hypotheses:

H1: There is a statistically significant relationship between the textual characteristics of a student's reflection and the presence of elaboration, reasoning, and linguistic richness in the reflection text.

H2: Changes in the content and structure of student reflections between initial and final drafts are positively associated with increases in elaboration, reasoning, and linguistic richness in the final text.

Objectives of the research work

This project aims to explore how student reflections can be analyzed using textual features to approximate indicators of reflective elaboration, reasoning, and linguistic richness – without relying on generative AI. The focus lies on data preprocessing and feature engineering using anonymized reflection data from the Reflect platform. The goal is to prepare a clean, structured dataset and define testable indicators that can support semi-automated assessment approaches in future work avoiding reliance on opaque generative AI systems (Zhang et al., 2024).

Structure of this research work

Chapter 2 provides a theoretical overview of relevant literature and concepts. Chapter 3 describes the methodological approach, including data preparation and feature extraction. Chapter 4 presents the results of the analysis. Chapter 5 discusses the findings, limitations, and implications. References and appendices conclude the work.

2. Literature Review / Theoretical Framework

This chapter outlines the theoretical background relevant to the automated analysis of student reflections. It focuses on reflection as a learning process, introduces established frameworks such as Gibbs' Reflective Cycle, and defines key textual and structural features that may indicate reflection depth or quality.

Reflection in Learning and Digital Contexts

Reflection is a core element of deep learning and personal development, allowing learners to make sense of their experiences, integrate new knowledge, and plan future actions (Moon, 1999). In educational contexts, reflective writing fosters

self-assessment, critical thinking, and personal meaning-making, which are essential in competency-based education frameworks like the IDG.

In technology-supported learning environments, reflection is often integrated into platforms or learning management systems (LMS) through structured prompts or writing tasks. However, assessing the quality or depth of such reflections remains challenging (Ullmann, 2019). Written reflections are typically open-ended and subjective, which makes scalable evaluation difficult — particularly when human judgment is impractical and reliance on opaque AI tools is undesirable.

This study addresses the issue by focusing on observable characteristics in student-written reflections that may act as proxies for reflective depth and development. Interpretable, text-based features offer a promising path toward transparent and reproducible assessment, especially when grounded in established models of reflective learning.

Gibbs' Reflective Cycle

One of the most widely used models in reflective writing is Gibbs' Reflective Cycle (1988), which proposes a structured process consisting of six stages: description, feelings, evaluation, analysis, conclusion, and action plan. Educational literature often emphasizes that reflections incorporating evaluation and analysis are deeper and more meaningful than those limited to mere description (Gibbs, 1988).

This model provides a useful conceptual foundation for identifying which aspects of a student's written reflection — such as cause-effect reasoning, learning statements, or goal setting — may signal more developed engagement with the reflection task. In this study, such indicators are approximated through features like reasoning keywords, word count, and sentence structure.

Moon's Levels of Reflective Writing

Complementing Gibbs' structural approach, Moon (1999) describes reflection in terms of levels of depth. At lower levels, reflective writing may be descriptive or narrative. At higher levels, it involves dialogic thinking, evaluation, and critical insight. This framework aligns well with attempts to quantify reflection depth based on observable textual characteristics, without requiring access to the learner's internal cognitive state.

Moon's model supports the assumption that increased elaboration, complexity, and use of reasoning language may indicate higher reflective engagement. This assumption is operationalized in this study through feature-based analysis. The approach taken here is most closely aligned with Ullmann's data-driven reflection models (Ullmann, 2015b; Ullmann, 2017), which rely on interpretable indicators such as reasoning expressions, linguistic complexity, and reflection structure – rather than mental models or black-box classifiers – to approximate reflection depth.

Key Concepts and Feature Proxies

To assess reflection without human or generative AI input, this study focuses on quantifiable text characteristics. These features are derived from written reflections and serve as indirect indicators of elaboration, reasoning, and richness. The main concepts used are:

Elaboration: Measured by features such as word count, sentence count, and average sentence length

Reasoning: Indicated by the presence of key terms like “because”, “learned”, or “understood” (see appendix 1c)

Linguistic richness: Captured through lexical diversity and readability indices (LIX/Läsbarhetsindex; Björnsson, 1968)

These concepts are not absolute measures of reflection quality but are treated as proxies that can be operationalized and tested statistically.

Derivation of the Research Hypotheses

Based on the theoretical models of Gibbs and Moon, and the operational definitions of elaboration, reasoning, and linguistic richness, this study derives its research focus from the assumption that these qualities are observable through structured text features. The hypotheses are grounded in the idea that such features can serve as proxies for reflection development, and that changes between initial and final drafts may indicate reflective growth. These assumptions are tested using anonymized student reflections, without relying on generative AI or manual evaluation.

Automated Assessment of Reflection Quality: State of Research

The automated assessment of student reflection quality has become an important topic in educational technology, driven by the need for scalable, transparent, and fair feedback. Early approaches relied on rule-based and keyword-matching techniques, identifying features such as reasoning keywords, word count, and lexical diversity as proxies for reflection depth (Ullmann, 2017; Birney, 2012). More recent work has explored ML and NLP-based approaches, leveraging n-grams, part-of-speech patterns, and tools such as LIWC to classify reflection quality and its indicators (Gibson et al., 2017; Ullmann, 2019).

Recent papers (e.g., the automated analysis in computer science reflective writing by Alrashidi et al., 2023) demonstrate that a combination of linguistic features, including first-person pronouns, causal connectors, and perspective-taking expressions, can reliably distinguish levels of reflection using classical NLP and ML models. These approaches have achieved moderate to substantial agreement with human raters, but are often preferred over black-box generative AI models for their transparency and explainability.

Overall, the literature confirms that interpretable, theory-driven textual features and interaction data can support the semi-automated, explainable assessment of reflection quality, particularly in multilingual and diverse educational contexts (Ullmann, 2019; Gibson et al., 2017).

Research Gap and Contribution

While prior research has established the feasibility of automated reflection assessment using linguistic features and ML, these methods are still rarely applied in practical, multilingual educational settings. This study extends existing approaches by implementing and validating transparent, rule-based text analysis methods for reflection quality on a real-world, multilingual dataset collected through the Rflect platform. The project thus contributes practical evidence on the use of scalable, explainable assessment without reliance on generative AI, particularly in the context of diverse student backgrounds and languages.

3. Methodology / How the Hypotheses are tested

3.1 Data acquisition

The dataset used in this study consists of 300 anonymized student reflections collected through the Reflect platform between Sept 30, 2024 and April 16, 2025. Each row captures a unique instance of the reflection process, combining input from the student, automated feedback from an AI model, and contextual metadata provided by the platform. The data is provided in an Excel file and includes a total of 13 attributes, covering textual responses, metadata from the reflection prompts, and behavioral indicators.

Following the import of the Excel file into a Python environment, an exploratory profiling analysis is conducted using ydata-profiling to assess data quality and structure. The dataset proves to be well-formed, with no duplicate entries and only minimal missing values, affecting approximately 3% of the records. These missing values are primarily found in non-critical fields such as *topic_description* and *reflection_snapshot*. The central analytical variable, *final_reflection*, is present in nearly all records. Some edge cases are identified, including zero-length reflections and negative values in *length_diff*, which are addressed in the data cleansing phase.

The attributes are categorized by their origin and function. From the lecturer, the dataset includes *topic_title*, *topic_description*, and *question*, which define the reflection prompt (details on the question catalogue see appendix 2). The student's initial response is stored in *reflection_snapshot*, which serves as the basis for AI-generated feedback captured in *suggestion_content* (for detailed overview see appendix 1). This feedback is generated using predefined parameters, recorded in *suggestion_params*. The student may then revise their text, producing the final reflection, which serves as the main input for subsequent analysis.

In addition, the dataset includes two timestamps, *created_at* and *last_reflection_moment*, to track the timing of the interaction. Quantitative measures such as *content_length*, *snapshot_length*, and *length_diff* describe the volume and change in text between drafts, while *seconds_spent* provides an estimate of the time the student engages with the reflection.

3.2 Data Integration

The dataset is reordered to ensure a consistent column structure that aligns with the analytical flow. Three additional fields are computed to standardize and validate the character-based content measures. The column *calc_refl_snap* records the actual number of characters in the field *reflection_snapshot*, and *calc_final_refl* reflects the character count of *final_reflection*. The length of the AI-generated feedback in *suggestion_content* is captured in *calc_ai_suggest_length*.

These fields are derived from the raw text and enable cross-verification with the original metadata fields *snapshot_length* and *content_length*. They also provide a consistent basis for subsequent analysis steps such as feature engineering and hypothesis testing.

3.3 Data cleansing

To ensure the uniqueness and analytical validity of the data, the dataset is first filtered to remove duplicate entries in the *final_reflection* field. Although only four entries are formally missing, the profiling report reveals that only 268 of 300 *final_reflection* texts are distinct. Upon review, 32 duplicate entries are identified and resolved by retaining only the most recent submission per unique text, based on the *last_reflection_moment* timestamp.

The *reflection_snapshot* field is also examined to assess the completeness and distinctiveness of initial drafts. A total of 56 entries lack a *reflection_snapshot*, representing instances where students did not submit an initial version of their reflection. These cases are retained for analyses focused solely on the final reflection (e.g., H1), but may be excluded from analyses that require both initial and final drafts (e.g., H2). No repeated entries remain in this field following the earlier deduplication process.

To ensure that only substantive reflections are included in the analysis, entries with a *final_reflection* length below 20 characters are removed after cleansing. This filtering step is applied to the HTML-stripped and whitespace-trimmed version of the text (*final_reflection_clean*) to reflect the actual content written by the student. In total, 4 entries are excluded through this criterion, resulting in a dataset of 264 records suitable for subsequent processing and analysis.

3.4. Feature Engineering

To support the quantitative analysis of reflective quality, the feature engineering process draws on the cleaned text field `final_reflection_clean`. Given the multilingual nature of the dataset, the process begins with language detection to ensure that subsequent analyses are linguistically appropriate.

Language Detection and Preprocessing

Each reflection is classified by its dominant language using the `langdetect` library. The dataset includes reflections written in English (122), German (103), French (22), and Dutch (15), with a few additional cases in minor languages. To ensure accurate sentence segmentation, the pipeline uses language-specific tokenizers from the NLTK `punkt` family. Each reflection is processed with the tokenizer that matches its detected language. This prevents mis-segmentation and preserves the syntactic structure necessary for reliable feature extraction.

Extraction of Linguistic Features

The following features are derived from each text to capture both structural and cognitive dimensions of reflective writing:

- **Text Length and Structural Complexity:** Word and sentence counts serve as indicators of elaboration and organization. Longer and more structured texts typically reflect more developed reasoning.
- **Lexical Diversity:** The ratio of unique words to total words quantifies vocabulary richness and expressive variation, both of which contribute to the depth of reflection.
- **Readability (LIX Index):** The LIX score accounts for sentence length and the presence of long words, providing a language-adjusted measure of syntactic complexity. The index is computed in a way that allows meaningful comparison across languages.
- **Reasoning Vocabulary:** A rule-based approach identifies the presence of keywords associated with reasoning and justification, based on the Inner Development Goals (IDG) framework. These keyword sets are manually translated and curated to preserve conceptual equivalence across English, German, French, and Dutch.

- **I-statements:** The frequency of sentence-initial personal pronouns (e.g., “I think,” “Ich glaube”) serves as a proxy for self-referential thinking, a key component of deeper personal reflection.

All features are computed using language-specific resources to account for linguistic variation. The approach is rule-based and interpretable, avoiding reliance on opaque, model-driven methods. This design supports transparency, reproducibility, and theoretical alignment with educational research.

Analytical Thresholds and Theoretical Grounding

To ensure analytical consistency and interpretability, several thresholds are applied to key features. These thresholds are informed by previous research on reflective depth and writing complexity (e.g., Kember et al., 2008; McCarthy & Jarvis, 2007; Björnsson, 1968), and align with frameworks such as data-driven NLP approaches proposed by Ullmann (2015b, 2017). They support both empirical classification and theoretical validity in assessing reflection quality.

Feature	Threshold	Justification	Source
Word Count	> 100 words	Indicates sufficient elaboration for structured reasoning	Kember et al., 2008
Lexical Diversity	> 0.4 (unique/total)	Indicates expressive, varied language, academic texts and reflective essays, values between 0.3–0.5 are typical.	Lu, 2012; McCarthy & Jarvis, 2007
Reasoning Keyword Count	≥ 2	Suggests presence of causal or justificatory thinking	Ullmann, 2017
Readability (LIX Index)	40–55	Reflects optimal sentence structure complexity, 20–30 = easy (e.g., children’s books) 40–55 = typical for essays, non-fiction, reflective texts	Björnsson, 1968

		>60 = overly complex, possibly unreadable	
--	--	--	--

Table 1: Summary of scoring criteria

Depth Scoring Logic

To provide a composite measure of reflection quality, a rule-based `depth_score` is computed by assigning one point for each of four criteria met by a reflection. The criteria are derived from linguistically interpretable features and grounded in prior research. Specifically, one point is assigned for each of the following: (1) word count exceeding 100 words, (2) lexical diversity greater than 0.4, (3) at least two reasoning-related keywords, and (4) a readability index (LIX) between 40 and 55. The total score therefore ranges from 0 to 4. This additive rubric enables interpretable classification of reflection depth and serves as the basis for the ordinal `depth_category` variable (Low, Moderate, High) used in subsequent statistical analysis.

3.5 Automation & Preprocessing Output

To support reproducibility and downstream analysis, the entire data preparation pipeline is implemented in Python and executed within a Jupyter Notebook environment. The final output is a fully cleaned, enriched dataset containing both original metadata and engineered linguistic features. Key steps in the automated workflow include:

- Removal of duplicates and incomplete entries.
- HTML stripping and whitespace normalization of all textual fields.
- Language detection and language-specific tokenization using NLTK models.
- Rule-based feature extraction for each final reflection, including reasoning, keyword detection, lexical diversity, sentence segmentation, and readability scoring.
- Calculation of derived metrics such as word differences and reasoning gains for before–after comparisons.

The output dataset contains 264 finalized entries and over 35 attributes, covering raw inputs, processed features, and derived indicators such as reflection depth score and revision behavior. The resulting file is exported in CSV format and serves as the basis for statistical hypothesis testing and visualization.

4. Results

4.1. H1: Relationship Between Linguistic Features and Reflection Depth

H1: “There is a statistically significant relationship between the textual characteristics of a student’s reflection and the presence of elaboration, reasoning, and linguistic richness in the reflection text..”

To investigate this hypothesis, the rule-based proxy for reflection depth (as defined in Section 3.4) classifies student reflections into three categories: Low Depth (score 0-1), Moderate Depth (score 2), and High Depth (score 3-4). For each engineered linguistic feature (e.g., word count, lexical diversity, reasoning indicators), the Kruskal–Wallis H test is applied to compare the distributions across the three depth categories. This non-parametric method is chosen due to the ordinal structure of the reflection depth variable and the skewed distribution of several linguistic metrics.

Statistical Results

The results show statistically significant differences ($p < 0.001$) for all features, supporting H1 (see Table 2). The structural features – word count, sentence count, and unique word count – increase markedly with reflection depth, indicating that deeper reflections are typically more elaborated and lexically varied. These trends are clearly visible in the boxplots, where the median and range expand consistently across depth levels. For instance, high-depth reflections contain substantially more words and sentences than low-depth ones, which often appear brief and structurally simple (see Figure 1).

...	# statistic	# p_value
word_count_final	156.5812699976399	9.97261918379712e-35
unique_words_final	154.75234738162462	2.4885979285545423e-34
sentence_count_final	136.25123339065127	2.5907206869009918e-30
lexical_diversity_final	125.357683615526	6.010697486084199e-28
reasoning_keywords_final	61.9759332870911	3.4841519035542403e-14
readability_lix_final	37.59362870968033	6.865106016233124e-09
i_statements_final	17.23443721172989	0.00018096288447680438

Table 2. Kruskal–Wallis Test Results for Linguistic Features by Depth Category

The analysis also reveals meaningful differences in reflective content. The frequency of reasoning-related keywords rises significantly with depth, suggesting that deeper reflections include more causal and evaluative reasoning. Likewise, I-statements, which signal self-referential thinking, appear more frequently in higher-depth reflections, although with greater variation. This supports the assumption that deeper reflection is not only more elaborate but also more personal and cognitively engaged (see Figure 2).

The findings for lexical diversity are less linear. While significant differences exist, lexical diversity appears to decline slightly in the highest-depth group. This may reflect increased repetition in longer texts, or a natural trade-off between elaboration and vocabulary range. Readability (LIX) also shows a significant distribution shift, although no consistent directional pattern emerges across categories. This suggests that syntactic complexity varies with depth but is likely influenced by language-specific factors.

Taken together, the results confirm H1: student reflections that score higher on a rule-based depth proxy also differ measurably in their linguistic structure and content. This demonstrates that reflective quality is aligned with concrete, interpretable text features – making automatic, rule-based analysis a viable and theoretically grounded approach.

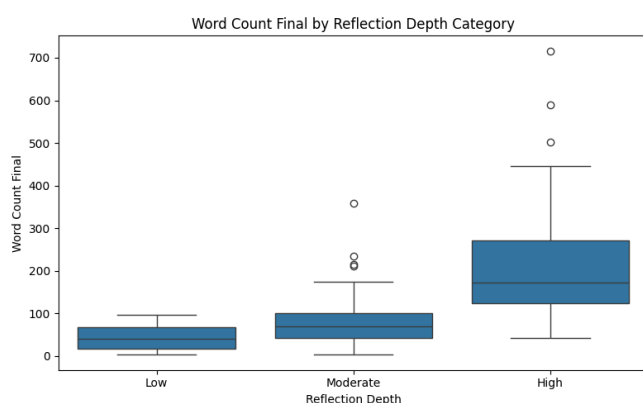


Figure 1: Distribution of Word Count Across Reflection Depth Categories (Low, Moderate, High)

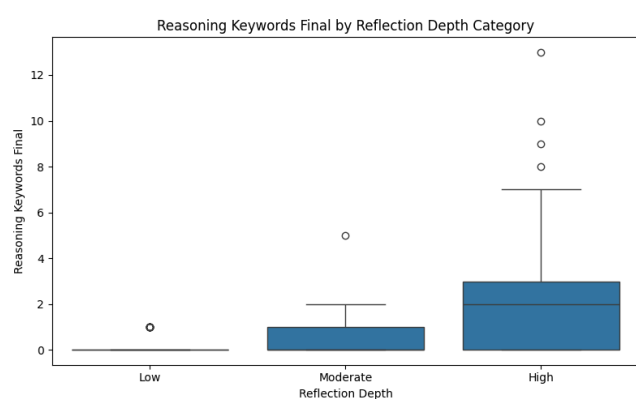


Figure 2. Frequency of Reasoning Keywords by Reflection Depth

4.2 H2: Correlation Between Revisions and Reflective Quality

Hypothesis 2 (H2) proposed that changes between the initial and final versions of student reflections are positively associated with higher levels of elaboration,

reasoning, and linguistic richness. To examine this hypothesis, a set of revision-based features was computed, and Spearman rank-order correlation analyses were performed.

The following change metrics are derived:

- `length_diff_clean` – character-level difference (cleaned) between final and snapshot text
- `word_diff` – change in total word count
- `reasoning_keyword_diff` – net increase in reasoning-related keywords
- `readability_diff` – change in readability index (LIX index)

These metrics are tested against two groups of dependent variables:

a) Final reflection quality features, including:

- word count, sentence count, and number of unique words
- lexical diversity and LIX readability
- counts of reasoning keywords and I-statements

b) The composite reflection depth score (`depth_score`), which integrates key indicators of elaboration, reasoning, and linguistic richness into a rule-based rubric.

Given the non-parametric distribution of the features, the analysis applies Spearman's ρ to assess correlation strength. The results are visualized using a heatmap and tabulated with p-values to evaluate statistical significance.

Results

The strongest observed association occurs between *reasoning_keyword_diff* and *reasoning_keywords_final* ($\rho = 0.52$, $p < 0.001$), indicating that students who introduce more reasoning-related terms during revision tend to produce reflections with greater cognitive engagement (see Figure 3).

Other revision metrics exhibit weak but statistically significant positive correlations with final quality indicators:

- *word_diff* and *unique_words_final*: $\rho = 0.29$, $p < 0.001$
- *word_diff* and *word_count_final*: $\rho = 0.28$, $p < 0.001$
- *length_diff_clean* and *unique_words_final*: $\rho = 0.28$, $p < 0.001$

No meaningful correlations are found for *readability_diff*, suggesting that changes in syntactic complexity do not consistently influence final quality outcomes.

When examining the relationship between revision metrics and the composite reflection depth score, the results show:

- *reasoning_keyword_diff*: $\rho = 0.274$, $p < 0.001$
- *word_diff*: $\rho = 0.251$, $p < 0.001$
- *length_diff_clean*: $\rho = 0.237$, $p < 0.001$
- *readability_diff*: $\rho = -0.025$, $p = 0.688$ (not significant)

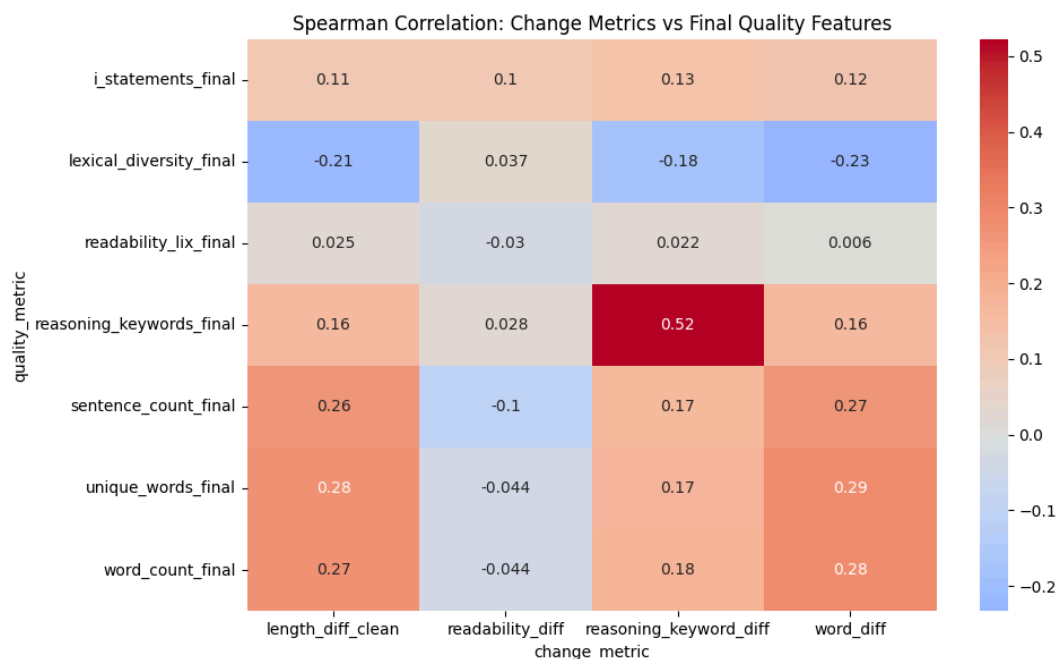


Fig 3: Spearman Correlation

These findings indicate that while revision activity – especially the inclusion of additional reasoning vocabulary – is positively associated with reflection depth, the strength of these relationships remains limited. In contrast, superficial or structural changes such as increased length or word count show only weak associations, and shifts in readability appear unrelated.

Conclusion on H2

The results provide partial support for Hypothesis 2. Revisions are positively associated with final reflection quality, particularly in terms of reasoning. However, the relatively modest correlation strengths suggest that the nature and quality of revision matter more than its quantity. These findings are consistent with

theoretical models of reflective writing, which emphasize meaningful cognitive elaboration over superficial textual expansion.

5. Discussion and Conclusions

This study investigates the extent to which students' reflection quality can be assessed automatically using linguistic features and revision behavior, without reliance on generative AI models. The research question guiding this investigation is: "How can students' reflection quality be automatically assessed using textual features and interaction data, without relying on generative AI?"

To address this question, the study applies a rule-based, multilingual feature engineering approach. Linguistic indicators are selected to serve as interpretable proxies for elaboration, reasoning, and linguistic richness. These features are extracted from both the final reflection text and the changes between initial and final drafts.

Summary of Findings

The results show that linguistic features are valid indicators of reflection depth (H1), and that meaningful revisions—particularly those involving reasoning—are positively associated with final quality (H2). The findings support the use of rule-based, language-aware metrics to assess reflection automatically.

Moreover, the results demonstrate that final reflection quality is not only measurable through structural features but also partially predictable from the nature and extent of revisions. In particular, conceptual expansion—rather than superficial textual growth—emerges as a more consistent indicator of deeper reflective engagement.

Answer to the Research Question

This study confirms that student reflection quality can be assessed through interpretable, rule-based linguistic and revision features, without relying on generative AI. The resulting method offers transparency, reproducibility, and multilingual applicability.

Educational and Practical Implications

The proposed approach provides a foundation for integrating automated reflection analytics into digital learning platforms. Lecturers can use these insights to deliver

formative feedback and promote deeper reflective practices. Emphasizing reasoning-focused revision, rather than superficial text expansion, appears to be a more effective instructional strategy. Furthermore, the study's use of transparent, rule-based, and language-sensitive methods promotes fair and interpretable assessment of reflection quality, especially in multilingual settings. This aligns with broader goals in educational technology, such as pedagogical transparency, learner agency, and equitable assessment – without relying on AI-based systems.

Limitations

This study relies on surface-level text features, which approximate but do not fully capture cognitive depth. Around 20% of reflections lack an initial draft, limiting the analysis of revision behavior. Additionally, rule-based keyword detection may miss subtle or culturally specific expressions of reasoning, particularly in a multilingual context. Given the multilingual nature of the dataset, variations in lexical richness and readability may reflect language-specific characteristics or inherent biases in linguistic feature extraction tools, which should be acknowledged when interpreting results.

Future Research Directions

Future work should explore hybrid models that combine rule-based and machine learning (ML) approaches. These could enhance classification accuracy while preserving interpretability. The inclusion of syntactic and discourse-level features may further improve depth estimation. Longitudinal designs and multi-turn reflections could also shed light on the development of reflective skills over time. Finally, equity-focused studies should examine the performance of such systems across diverse linguistic and cultural groups.

Conclusion

The findings demonstrate that rule-based textual features and revision metrics can serve as reliable proxies for reflective quality. This approach offers a transparent, interpretable alternative to black-box models and contributes to the development of explainable learning analytics in reflective education. By focusing on human-understandable indicators such as elaboration, reasoning vocabulary, and linguistic richness, the study supports a scalable yet pedagogically grounded method for assessing reflection quality.

References

- Alrashidi, H., Almujaally, N., Kadhum, M., Ullmann, T. D., & Joy, M. (2023). Evaluating an automated analysis using machine learning and natural language processing approaches to classify computer science students' reflective writing. In G. Ranganathan, R. Bestak, & X. Fernando (Eds.), *Pervasive Computing and Social Networking* (Vol. 475, pp. 401–415). Springer. https://doi.org/10.1007/978-981-19-2840-6_36
- Björnsson, C. H. (1968). *Läsbarhet* (Readability). Stockholm: Liber.
- Cendon, E. (2016). Die Rolle der Reflexion in der Weiterbildung. In R. Egger & M. Merkt (Hrsg.), *Lernwelt Erwachsenenbildung* (S. 241–257). Springer VS. https://doi.org/10.1007/978-3-658-11691-0_14
- Gibbs, G. (1988). *Learning by doing: A guide to teaching and learning methods*. Oxford Polytechnic, Further Education Unit. <https://trove.nla.gov.au/work/16621092>
- Gibson, A., Aitken, A., Sándor, Á., Buckingham Shum, S., Tsingos-Lucas, C., & Knight, S. (2017). Reflective writing analytics for actionable feedback. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK '17)* (pp. 153–162). Association for Computing Machinery. <https://doi.org/10.1145/3027385.3027436>
- Inner Development Goals Initiative. (2021). *Inner Development Goals: A framework for personal and collective growth*. <https://www.innerdevelopmentgoals.org/>
- Kember, D., McKay, J., Sinclair, K., & Wong, F. K. Y. (2008). A four-category scheme for coding and assessing the level of reflection in written work. *Assessment & Evaluation in Higher Education*, 33(4), 369–379. <https://doi.org/10.1080/02602930701293355>
- Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96(2), 190–208.
- McCarthy, P. M., & Jarvis, S. (2007). Vocab: A theoretical and empirical evaluation. *Language Testing*, 24(4), 489–496. <https://doi.org/10.1177/0265532207080767>
- Moon, J. A. (1999). *Reflection in learning and professional development: Theory and practice* (1st ed.). Routledge. <https://doi.org/10.4324/9780203822296>
- Moon, J. A. (2004). *A handbook of reflective and experiential learning: Theory and practice*. RoutledgeFalmer. <https://books.google.ch/books/about/>

[A_Handbook_of_Reflective_and_Experientia.html?id=vs5dJozQSdwC&redir_esc=y](#)

- Moon, J. A. (2006). *Learning journals: A handbook for reflective practice and professional development* (2nd ed.). Routledge.
<https://doi.org/10.4324/9780203969212>
- Ryan, M. (2011). The pedagogical balancing act: Teaching reflection in higher education. *Teaching in Higher Education*, 16(1), 99–110.
<https://eprints.qut.edu.au/218804/1/54644.pdf>
- Schön, D. A. (1992). *The reflective practitioner: How professionals think in action* (1st ed.). Routledge. <https://doi.org/10.4324/9781315237473>
- Ullmann, T. D. (2017). Reflective writing analytics: Empirically determined keywords of written reflection. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK '17)* (pp. 163–167). Association for Computing Machinery. <https://doi.org/10.1145/3027385.3027394>
- Ullmann, T. D. (2019). Automated analysis of reflection in writing: Validating machine learning approaches. *International Journal of Artificial Intelligence in Education*, 29(2), 217–257. <https://doi.org/10.1007/s40593-019-00174-2>
- Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, 100, 94–109.
<https://doi.org/10.1016/j.compedu.2016.05.004>
- Wong, F. K., Kember, D., Chung, L. Y. F., & Yan, L. (1995). Assessing the level of student reflection from reflective journals. *Journal of Advanced Nursing*, 22(1), 48–57. <https://doi.org/10.1046/j.1365-2648.1995.22010048.x>
- Zhang, C., Hofmann, F., & Plöchl, L. (2024). Classification of reflective writing: A comparative analysis with shallow machine learning and pre-trained language models. *Education and Information Technologies*, 29, 21593–21619.
<https://doi.org/10.1007/s10639-024-12720-0>

Appendix

The project is on GitHub at:

<https://github.com/pluzgi/FDA93SPIII4>

Accompanying this thesis, the following files were created:

- Jupyter Notebook with Python code (“main.ipynb”)
- Declaration of independent work
(“250607_Eigenständigkeitserklärung_SabineWildemann_FDA93SPIII4.pdf”)
- Presentation (“250607_Presentation_FDA93SPIII4_SabineWildemann.pdf”)
- Screencast ([Teams link](#)) (“LINK”)
- Zip file (“250607_FDA93SPIII4_Prüfungsleistung_SabineWildemann.zip”)

Data set

Excel file, “dataset_rflect.xlsx”, 313 kB

1. a) Original dataset attributes – Source-based overview

Source / Type	Attribute(s)
from lecturer	“topic_title”, “topic_description”, “question”
from student	“reflection_snapshot” (initial version of the reflection)
from AI model	“suggestion_content” (referring to the reflection_snapshot)
from student	“final_reflection” (after receiving AI model suggestion(s))
timestamps (date, time):	“created_at”, “last_reflection_moment”
predefined model parameters (like model type etc.)	“suggestion_params”
amount of characters of students initial “reflection_snapshot”	“snapshot_length”
amount of characters of students “final_reflection”	“content_length”

difference between "reflection_snapshot" /"final_reflection"	"length_diff"
duration of time spent for reflection	"seconds_spent"

b) Feature-engineered attributes – Created for analysis and modeling

'topic_title', 'topic_description', 'question', 'reflection_snapshot', 'snapshot_length',
'calc_refl_snap', 'suggestion_content', 'calc_ai_suggest_length', 'final_reflection',
'content_length', 'calc_final_refl', 'length_diff', 'created_at',
'last_reflection_moment', 'seconds_spent', 'suggestion_params', 'topic_title_clean',
'topic_description_clean', 'question_clean', 'reflection_snapshot_clean',
'suggestion_content_clean', 'final_reflection_clean', 'final_length_clean',
'snapshot_length_clean', 'calc_ai_suggest_length_clean', 'length_diff_clean',
'lang_detected', 'word_count_final', 'sentence_count_final', 'unique_words_final',
'lexical_diversity_final', 'readability_lix_final', 'reasoning_keywords_final',
'reasoning_phrases_final', 'i_statements_final', 'depth_score', 'depth_category'

c) Overview of reasoning keywords and reasoning phrases used

"en"

"because", "learned", "learning", "understood", "realized", "noticed",
"aware", "reflect", "reason", "consider", "analyze", "evaluate",
"growth", "development", "perspective", "viewpoint", "insight",
"therefore", "thus", "hence", "however", "although", "nevertheless",
"moreover", "furthermore", "alternatively", "in contrast", "as a result",
"implies", "suggests", "indicates"

"de"

"weil", "gelernt", "lernen", "verstanden", "bemerkt", "realisiert",
"bewusst", "reflektieren", "grund", "überlegen", "analysieren", "bewerten",
"wachstum", "entwicklung", "perspektive", "sichtweise", "einsicht",
"deshalb", "somit", "daher", "jedoch", "obwohl", "nichtsdestotrotz",
"außerdem", "ausserdem", "weiterhin", "alternativ", "im gegensatz", "als
ergebnis",
"impliziert", "deutet hin", "zeigt"

"nl":

"omdat", "geleerd", "leren", "begrepen", "beseft", "opgemerkt",
"bewust", "reflecteren", "reden", "overwegen", "analyseren", "evalueren",
"groei", "ontwikkeling", "perspectief", "inzichten", "standpunt",
"daarom", "dus", "vandaar", "echter", "hoewel", "desondanks",
"bovendien", "verder", "alternatief", "daarentegen", "als resultaat",
"suggereert", "wijst op", "duidt aan"

"fr":

"parce que", "appris", "apprendre", "compris", "remarqué", "réalisé",
"conscient", "réfléchir", "raison", "considérer", "analyser", "évaluer",
"croissance", "développement", "perspective", "point de vue", "aperçu",
"donc", "ainsi", "par conséquent", "cependant", "bien que", "néanmoins",
"de plus", "en outre", "alternativement", "en revanche", "en résultat",
"suggère", "indique", "implique"

"en": "on the other hand", "i think", "i believe", "i realize", "i notice", "i conclude"

"de": "andererseits", "ich denke", "ich glaube", "ich erkenne", "ich bemerke", "ich
schliesse", "ich schließe"

"nl": "aan de andere kant", "ik denk", "ik geloof", "ik realiseer me", "ik merk op", "ik
concludeer"


"fr": "d'autre part", "je pense", "je crois", "je réalise", "je remarque", "je conclus"

2. Reflect Question catalogue, oriented on IDG


Foundational Topics (Program/Process-Oriented)	Self-assessment Debrief: Program Start Program Start: Expectations & Learning Goals Reflecting on Your First Experiences How you Reflect Module Preparation Post-Module Reflection: Key Takeaways Post-Module Reflection: Insights and Next Steps Check-in Insight to Action: Bridging the Gap From Insight to Action: Making It Work Group Work: Reflecting on Your Role Getting Ready for Exams
---	--

	Self-assessment Debrief: End of Program Closing Reflection
IDG: Being (Self-Development)	Inner Compass Exercise: Living your Values Integrity & Authenticity Openness & Learning Mindset Self-awareness Presence
IDG: Thinking (Cognitive Skills)	Critical Thinking Complexity Awareness Perspective Skills Sense-making Long-term Orientation & Visioning
IDG: Relating (Emotional & Social Awareness)	Appreciation Connectedness Humility Empathy & Compassion
IDG: Collaborating (Team & Group Dynamics)	Communication Giving and Receiving Feedback Interpersonal Relationships Co-creation Inclusive Mindset & Intercultural Competence Trust Mobilisation
IDG: Acting (Action & Resilience)	Courage Creativity Optimism Perseverance Resilience Celebration

3. Example of Self-Reflection task, IDG: “Relating”


 Reflect

100%



Lecturer Preview Draft - Not visible to students

Humility



"True humility is not thinking less of yourself; it is thinking of yourself less." – C.S. Lewis Humility shows up in collaboration, when you set aside ego for the collective good. It is also evident when you admit mistakes and learn from them.

1. Describe a situation where you initially believed you were right about something, but later realized you were mistaken. How did you handle admitting your mistake, and what did this teach you about humility?

0 Words

2. How does acknowledging your limitations open new opportunities?

0 Words

3. How can you practice humility in your daily interactions?

0 Words