GENIUS Keylog Corpus - A German High School Student Corpus with Keystroke Logging Data

Nils-Jonathan Schaller[1], Thorben Jansen[1], Lars Höft[1], Hannah Pünjer[1], Andrea Horbach[1,2]

schaller@leibniz-ipn.de
[1]Leibniz Institute for Science and Mathematics Education, Germany, [2]Kiel University, Germany
schaller@leibniz-ipn.de

Abstract

We present the GENIUS Keylog corpus: 259 German high school student essays with keystroke logging data and comprehensive argumentative annotations. The corpus enables temporal alignment between writing processes and linguistic structure—showing when and how argumentative elements were produced. We provide corpus statistics, methodologies for linking keystroke data to linguistic annotations, and first analyses of the dataset. The corpus enables the investigation of relationships between writing processes and argumentative structure, supporting research on process-aware feedback systems and writing under time pressure.

Keywords: Keystroke logging, corpus, argument mining

1. Introduction

Understanding student writing requires examining both the final text (product) and how it was composed (process). Existing research has studied these dimensions separately: argumentation corpora such as DARIUS (Schaller et al., 2024) PERSUADE (Crossley et al., 2022) provide rich linguistic annotations of argumentative structure, while keystroke logging studies (Conijn et al., 2025; Schiller et al., 2025) capture writing processes through measures such as pauses, revisions, and fluency patterns. However, to the best of our knowledge, no work has combined keystroke logging data with fine-grained, linguistic annotations of argumentative structures. This gap limits our understanding: Do students revise different parts of arguments differently? Does writing speed vary across essay sections? Do pauses cluster around specific structural elements? Aligned process and product data would bridge the gap between information on how students write and what they write, providing insights into the underlying processes of argumentative composition and enabling process-aware feedback on writing.

We present the GENIUS (Generative Intelligence Utilized in Schools) Keylog corpus: to the best of our knowledge, the first publicly available corpus combining keystroke logging data with comprehensive argumentative annotations. While corpora such as KLiCKe (Tian et al., 2025) provide keystroke data with holistic quality scores, GENIUS enables temporal alignment between specific argumentative elements and their production process, as illustrated in Figure 1 for an example essay.

Our main contributions are the following: We release the GENIUS Keylog corpus with 259 German argumentative essays on socio-scientific issues in relation to climate change, accompanied by keystroke logging data with a total of 233,719 events and comprehensive argumentative annotations, provided as separate, linkable data files. We provide baseline analyses of zone-specific writing behaviors under time pressure, revealing systematic patterns in revision, speed, and time allocation across essay sections.

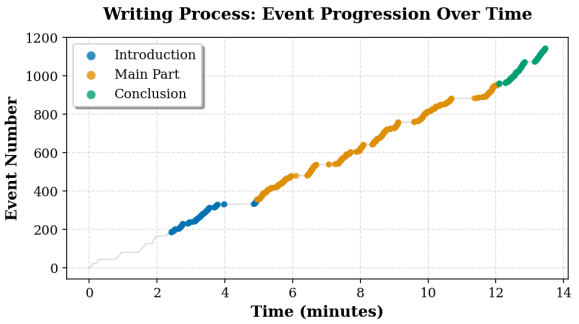

Figure 1: Figure 1: Example essay writing process (essay id: 1008)

We establish methodologies for linking keystroke data to linguistic structure, enabling future research on relationships between writing processes and argumentative quality.

All data, annotations, and analysis code are publicly available at https://github.com/darius-ipn/genius_keylogs.git.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6919–6928
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2. Related Work

This section reviews prior work in three areas: keystroke logging datasets for writing research, process-based feedback systems and writer profiling, and argumentation corpora. This review establishes how GENIUS fills a gap at the intersection of these research streams.

2.1. Keystroke Logging Datasets for Writing Research

Several corpora provide keystroke logging data for writing research, though they differ in scale, annotation depth, and language coverage.

Velentzas et al. (2024) released the KUPA-KEYS dataset with keystroke logs from 1,006 participants who had written essays in English, accompanied by CEFR proficiency assessments. While this dataset enables correlation analysis between keystroke behaviors and overall writing quality, it provides only holistic scores without fine-grained linguistic annotations.

The Klicke corpus by Tian et al. (2025) comprises approximately 5,000 essays written by US adults recruited through MTurk, composed in 30-minute sessions in response to prompts adapted from retired SAT assessments. It provides comprehensive process features (pauses, bursts, revisions) and, like KUPA-KEYS, holistic quality scores, but it lacks structural annotations of argumentative content.

In contrast, Mahlow (2014) combined keystroke logging with sentence-level morphosyntactic annotations for German, providing methodologies for linking process data to linguistic structure. However, this work focused on syntactic rather than argumentative annotations.

These datasets enable the investigation of relationships between writing processes and overall quality or linguistic features. However, to the best of our knowledge, no existing corpus combines keystroke logging data with detailed annotations of argumentative structure, which is necessary to investigate questions such as how revision patterns differ across argument components or how planning pauses cluster around specific structural elements.

2.2. Process-Based Feedback Systems and Writer Profiling

Keystroke logging has been employed to develop feedback systems that target writing processes and to identify distinct writer profiles for personalized instruction.

Vandermeulen et al. (2024) demonstrated that keystroke logging-based process feedback significantly improved writing among Dutch students. However, their analysis focused exclusively on behavioral measures without examining the linguistic or argumentative content being produced, leaving it unclear how these behaviors relate to argumentative structure and quality.

In our previous work (Schaller et al., 2025), we showed that AM models trained on complete essays suffer substantial performance drops on artificially fragmented texts, particularly for context-dependent elements like conclusions. The GENIUS Keylog corpus addresses this limitation by providing authentic writing process data rather than artificial truncations.

Schiller et al. (2025) investigated behavioral engagement in writing revision using keystroke logging with 453 EFL learners, measuring metrics including total keystrokes, typing time, and writing pauses to understand how engagement mediates automated feedback effectiveness. Their findings suggest that process-based feedback should account for individual engagement patterns.

Mahlow (2015) applied a process-product framework to error analysis, demonstrating how keystroke logging data can reveal that apparent surface errors often result from complex revision operations rather than simple typos. This work highlights the diagnostic potential of keystroke data for targeted feedback.

Building on the feedback perspective, several studies have used keystroke logging to identify writer profiles that could inform differentiated instruction. Zhang et al. (2019) applied clustering techniques to keystroke data from 740 middle school students who had written argumentative essays and identified four distinct profiles ranging from fluent to struggling writers. Their analysis of sequential patterns revealed how editing behaviors and fluency evolved differently across profiles. Similarly, Talebinamvar and Zarrabi (2022) identified five distinct writer profiles in EFL argumentative writing using keystroke feature clustering and demonstrated that under less constrained conditions, writer profiles reflect qualitative strategic differences beyond mere productivity.

While these studies demonstrate the value of keystroke logging for process-aware feedback and writer profiling, they analyze process patterns independently of the specific argumentative content being produced. This limits the development of feedback systems that could target both process behaviors and argumentative quality simultaneously.

3. GENIUS Keylog Corpus

The GENIUS Keylog corpus builds on the design of the DARIUS corpus by replicating its argumentative writing task, but it constitutes an independent data collection. It includes keystroke logging data

from 273 students at German high schools who wrote essays discussing propulsion technologies for automobiles, such as battery-electric vehicles, hydrogen fuel cells, and synthetic e-fuels, and evaluated their advantages and disadvantages in the context of climate change. Students were given 15 minutes to write their essay.

Participant characteristics are described in Table 2: The sample shows variation in age (range: 14-19 years, M = 16.36, SD = 0.74) and gender (60.6% female, 32.8% male, 0.8% non-binary), while household language is more homogeneous, with 86.1% reporting German as their household language. All participants were in grades 10 and 11.

3.1. Keystroke Logging

For each essay, the system recorded every keystroke event, including timestamps (milliseconds) and information about which key was pressed. This process data enables the reconstruction of the complete writing timeline, including text production, deletion, and pauses. From 273 texts with keystroke data collected, 259 were suitable for analysis after files that were either too short (less than a sentence) or off-topic or had data protection issues were removed. Students spent an average of 5.23 minutes actively writing and were engaged for an average of 11.16 minutes on the total task until the last keystroke, generating a mean of 1,034 keystrokes per essay.

An overview of selected keylog statistics is displayed in Table 1.

Characteristic	**Value**
Total participants	259
Average total time	542 sec (9 min)
Average total events	931
Writing Strategy	
Linear writers	249 (96%)
Non-linear writers	10 (4%)
Average zone switches	1
Writing Bursts	
Average number of bursts	28
Average burst duration	11 sec
Fluent bursts	22 (79%)
Revision bursts	2 (9%)
Mixed bursts	3 (12%)
Pause Behavior	
Short pauses (mean)	697 ms
Mid pauses (mean)	1.358 ms
Normal pauses (mean)	3.079 ms
Long pauses (mean)	13.450 ms
Longest pause	28 sec

Table 1: Overall Writing Process Characteristics

Variable	*n*	**%**
Age		
Mean (SD)	16.36 (0.74)	-
Range	14–19	-
Missing	2	0.8
Gender		
Female	157	60.6
Male	85	32.8
Non-binary	2	0.8
Missing / No answer	15	5.8
Household language		
German	223	86.1
Russian	6	2.3
Other	30	11.6
Missing	4	1.5

Table 2: Study Population Characteristics (N = 259)

3.2. Annotation Scheme

The GENIUS Keylog corpus includes annotation layers adapted from the DARIUS annotation scheme (Schaller et al., 2024), though this paper focuses exclusively on content zone annotations in order to link them to keystroke data as the basis for analyzing zone-specific writing processes (Section 4), as they represent distinct functional units with high inter-annotator agreement. We briefly describe all annotation layers, as they are available in the corpus for future research. Six annotators were trained and each essay was annotated by at least two separate annotators.

3.2.1. Content Zones

All 259 essays were annotated for content zones and the three structural sections of argumentative essays, following the guidelines for content areas described by Stede et al. (2015): *introduction*, *main part*, and *conclusion*.

Cohen's κ for content zone boundaries was 0.83, indicating substantial agreement. Content zones provide a reliable structural framework with clear functional distinctions:

- **Introduction:** Introduces the topic, provides context, establishes the writer's position.
- **Main Part:** Presents arguments, evidence, and reasoning.
- **Conclusion:** Synthesizes arguments and concludes.

3.2.2. Additional Annotation Layers

Following DARIUS, the corpus includes three additional hierarchical annotation layers:

- **Arguments:** Individual argumentative units within the *main part*, annotated for topic, position (support/oppose), and quality dimensions (accuracy, adequacy).
- **Major Claims:** The writer's overall position (Stab and Gurevych, 2014) and associated decision-making strategy.
- **Toulmin Elements:** Argumentation components (claim, data, warrant, rebuttal) following Toulmin's argumentation pattern by Riemeier et al. (2012).

These layers are available in the released corpus and enable future research on relationships between writing processes and argumentative structure.

4. Analysis of Writing Processes in Argumentative Essays

Before presenting our analyses, we define the key metrics derived from keystroke logs and explain our methodology for linking them to content zones.

Basic Metrics For every key pressed, we extracted the timestamp (in milliseconds) and the character offset. This allowed us to reconstruct the complete writing process, including text production, deletion, and pauses. We categorized keystroke events as either *productive* (character insertions and line breaks) or *deletions* (backspace operations). The *revision ratio* for a content zone is calculated as the number of deletions divided by the number of productive events within that zone.

Pause Classification While classic keystroke logging studies often define a pause as an inter-keystroke interval of 2 seconds or more (Wengelin, 2006), we extended our analysis to include a wider range of pause durations to capture more nuanced typing behaviors:

- *None*: $<$ 500ms (typing flow, no meaningful pause)
- *Short pauses*: 500–999ms
- *Mid pauses*: 1.000–1.999ms
- *Normal pauses*: 2.000–4.999ms
- *Long pauses*: $\geq$ 5.000ms

Writing Bursts Since a writing burst refers to uninterrupted text production (Leijten et al., 2019), the definition of a pause determines the definition of a burst. We define *writing burst* as a continuous sequence of at least 5 keystroke events that can be separated by pauses of less than 2 seconds. Bursts are classified based on their revision ratio: *fluent* bursts have a revision ratio $<$ 0.2, *revision-heavy* bursts have ratios $>$ 0.5, and *mixed* bursts fall between these thresholds.

Productivity Metrics Beyond raw event counts, we calculated *productive characters* for each content zone by subtracting deletions from total events (productive characters = events - deletions). We computed writing speed as characters per second.

Linking Keystroke Data to Content Zones As only the final state of an essay was annotated, a process to match keystroke events with annotations was needed. The system reconstructs the state of the text at any keystroke event. Annotated tokens are grouped into sentences based on ending punctuation, with each sentence assigned the most common label among its tokens. For each keystroke event, the last sentence present in the text at that moment in the current text snapshot is extracted and compared against all annotated sentences using Python's difflib.SequenceMatcher[1], which calculates string similarity based on the longest common substring. The algorithm performs case-insensitive comparisons and returns similarity scores between 0.0 and 1.0. A threshold of 0.6 (60% similarity), which is recommended in the library's documentation, is required to accept a match. When a match is found, the keystroke event inherits the content zone label from the matched annotated sentence. Matches are cached to avoid redundant comparisons for consecutive keystrokes within the same sentence. This fuzzy matching approach accommodates typos, incomplete sentences, and minor edits between the writing process and final annotations, though heavily revised content or deletions may fail to match and remain labeled "unknown."

4.1. Task Completion under Time Constraints

Task completion patterns reveal how students prioritize and manage limited time. Only 12.0% of students (n = 31) included all three sections of an argumentative essay in their text. The majority, 88.0% (n = 228), did not write a *conclusion* (see Table 3).

The breakdown of essay structures reveals several patterns. Most commonly (60.2%, n = 156), students wrote an *introduction* and *main part* but no *conclusion*. Another 14.3% (n = 37) wrote only an introduction, never progressing to the main part. The remaining students either completed all three sections (12.0%) or followed other patterns such as writing only the *main part* (13.5%). Most of

[1] https://tedboy.github.io/python_stdlib/_modules/difflib.html

Completion Pattern	n	%
Completed all three sections	31	12.0
No conclusion written	228	88.0
Only *introduction*	37	14.3
introduction + main part	156	60.2
Other patterns	35	13.5
Total	**259**	**100.0**

Table 3: Essay Completion Rates

these observations suggest a predominantly linear composition: Students appear to write sections sequentially rather than switching between them.

Time allocation data further illuminate these patterns. Table 4 shows that students divided their time approximately equally between the *introduction* (M = 46.9%, SD = 31.5%) and *main part* (M = 50.2%, SD = 31.1%) and spent only 2.6% of their writing time on conclusions.

Content Zone	Mean (%)	SD	Median (%)
introduction	46.9	31.5	41.6
main part	50.2	31.1	55.8
conclusion	2.6	11.1	0.0

Table 4: Time Allocation Across Content Zones

These patterns can be interpreted in multiple ways. Students might prioritize developing their arguments over crafting formal conclusions. Alternatively, it may indicate that students underestimate the time needed for complete essays or lack strategies for allocating writing time to all three sections within 15 minutes.

4.2. Writing Strategies: Linear Progression under Time Pressure

Among the students who did work across all three content zones, writing strategies were very uniform. Of the 259 students, 96.1% (n = 249) demonstrated linear writing patterns, progressing sequentially through sections without recursion. These students made forward transitions between zones (e.g., *introduction* $\rightarrow$ *main part*) but never returned to a previously visited zone.

Non-linear writers (3.9%, n = 10) returned to previously visited zones at least once.

The dominance of linear writing is unsurprising given the 15-minute time constraint. Under time pressure, forward-moving might be more efficient than switching between sections or revising.

4.3. Time Allocation between Content Zones

Given that 88% of students wrote no *conclusion*, our time allocation analysis primarily concerned the distribution of effort between the *introduction* and *main part*. Among the students who worked on all three sections, 56.0% (n = 145) spent the majority of their time on the *main part*, while 39.4% (n = 102) focused primarily on the *introduction*. This nearly even split masks two qualitatively different writing patterns.

Students who spent more time on the *main part* likely represent the 60.2% who completed both the *introduction* and *main part* but stopped before the *conclusion*. In contrast, students who spent most time on the *introduction* likely overlap substantially with the 14.3% who wrote only an *introduction* and never transitioned to the main part.

To control for text length, we compared the proportion of text produced in each section with the amount of time spent there. For each student, we calculated a ratio of text proportion to time proportion. A ratio of 1.0 indicates proportional output (e.g., 50% of the time produced 50% of the text). Ratios below 1.0 mean less text relative to time (lower efficiency), ratios above 1.0 mean more text relative to time (higher efficiency).

Students were less efficient in the *introduction* (0.91) than in the *main part* (1.06), which means they spent more time per unit of text on the introduction.

4.4. Zone-Specific Revision Patterns

Revision behavior, measured as the ratio of deletions to productive keystrokes, varied systematically across content zones (see Table 5). The *introduction* showed the highest revision ratio (M = 0.143), followed by the *main part* (M = 0.106). The *conclusion* showed minimal revision (M = 0.010), but only 12.0% of students actually wrote conclusions.

Content Zone	Mean	SD	n
introduction	0.143	0.107	224
main part	0.106	0.082	209
conclusion	0.010	0.034	31
Overall	0.145	0.082	259

Table 5: Revision Ratios by Content Zone

The higher revision ratio in introductions compared to main parts may reflect the challenge of finding an engaging starting point for the essay. As the writers wrote mostly linearly, it was in the *introduction* that they already decided the position they wanted to argue, which may require more

trial-and-error and revision. In contrast, once the *introduction* is established, generating supporting points and evidence in the *main part* may proceed more fluently with less need for revision.

4.5. Writing Speed and Fluency

Writing speed varied between the two zones that most students actually composed (see Table 6). Students typed fastest in the *main part* (M = 1.64 chars/second) and slightly slower in the *introduction* (M = 1.52 chars/second). The reported *conclusion* writing speed (M = 0.36 chars/second) should be interpreted with caution, given that only 12.0% of students wrote a *conclusion* at all.

Content Zone	Mean (chars/sec)	SD
introduction	1.52	0.73
main part	1.64	0.91
conclusion	0.36	1.05

Table 6: Writing Speed by Content Zone

The faster writing speed in the *main part* compared to the *introduction* may indicate that once students establish their argumentative framework in the introduction, they can generate supporting content more fluently.

The analysis of writing bursts revealed that students engaged in an average of 27.5 bursts per essay, with a mean duration of 10.5 seconds. The majority of these bursts were fluent (78.9%), characterized by low revision ratios, while only 8.9% were revision-heavy bursts, and 12.2% showed mixed patterns (see Table 7).

Burst Type	Mean per Essay	% of Total
Fluent	21.7	78.9
Revision-heavy	2.4	8.9
Mixed	3.3	12.2
Total bursts	**27.5**	**100.0**

Table 7: Writing Burst Patterns

The prevalence of fluent bursts suggests that most students' writing processes are dominated by periods of relatively uninterrupted text generation rather than extensive revisions. This fluent, forward-moving pattern aligns with the linear writing strategies we observed.

4.6. Pause Behavior and Planning

Pauses might indicate planning or reading. The vast majority of pauses represented continuous typing flow with no meaningful pause. Short pauses comprised 11.4%, and mid-length pauses accounted for 5.0%. Longer pauses, suggestive of cognitive processing or distraction, were relatively rare: Normal pauses comprised only 2.4% of all pauses, while long pauses accounted for just 1.9% (see Table 8).

The average longest pause per student was 28.2 seconds, amounting to just 3.1% of the 15-minute writing period. This indicates that almost all students avoided longer planning or reflection, likely due to time constraints. This pattern again suggests a focus on continuous writing and fluency, with even the longest pauses likely used for reading or immediate planning, rather than revision.

Pause Category	Duration	Mean %
None (typing flow)	$<$ 500ms	79.2
Short pauses	500–999ms	11.4
Mid pauses	1000–1999ms	5.0
Normal pauses	2000–4999ms	2.4
Long pauses	$\geq$ 5000ms	1.9

Table 8: Pause Behavior Distribution

4.7. Correlation Between Grades and Writing Behavior

Students self-reported their most recent grades in six subjects (German, English, math, chemistry, physics, and biology) on a scale from 1 (very good) to 6 (unsatisfactory) following the German education system. An exploratory analysis revealed limited relationships between writing behaviors and grades. Only mathematics showed a significant association with *conclusion* completion (students who wrote conclusions: M = 2.13, SD = 0.97; students who did not: M = 2.59, SD = 1.04; t = -2.25, p = .026, Cohen's d = -0.44). Physics grades showed weak negative correlations with both output (r = -.197, p = .003) and speed (r = -.160, p = .019), and chemistry showed a weak correlation with output (r = -.157, p = .021). In the multiple comparisons conducted (18 tests across 6 subjects and 3 metrics), most relationships were non-significant, suggesting that writing behavior in this constrained 15-minute task may not strongly relate to broader academic performance. Detailed results are available in the appendix: Tables 9, 10, and 11.

5. Exploratory Analysis: Writer Profiles

Section 4 examined zone-specific writing behaviors to understand how students compose different parts of argumentative essays. This structural lens revealed systematic patterns in revision, speed, and time allocation across essay sections. However, writing research also seeks to identify writer

profiles, e.g., characteristic patterns that distinguish individual writers from each other regardless of which part of the essay they are working on. cite To complement the zone-specific analysis, we conducted an exploratory investigation of whether students cluster into distinct writer types based on their overall keystroke patterns. While Section 4 examined whether behaviours vary across zones, this analysis focuses on how writers vary from each other, using essay-level features aggregated across entire texts.

5.1. Method

We extracted 47 keystroke features from each essay, capturing overall session metrics (total time, keystrokes, text length), burst patterns (continuous typing sequences), pause behavior (duration and frequency), and deletion patterns. All features were z-score normalized to ensure their equal contribution to clustering. (*Note that, unlike for Zhang et al. (2019) and Talebinamvar and Zarrabi (2022), no holistic quality scores were available to us.*)

A Principal Component Analysis (PCA) reduced the feature space from 47 to 11 components, explaining 81.5% of the variance. The first component (PC1, 22.7% of the variance) reflected overall productivity: total keystrokes (loading: 0.931), pauses (0.931), and active writing time (0.869). The second component (PC2, 11.9% of the variance) captured editing behavior: mean burst length (0.785), discarded text ratio (0.713), and deletion metrics (0.647).

We applied K-means clustering to the 11 PCA components. The optimal cluster number was determined using silhouette analysis, which favored $K = 2$ (silhouette score: 0.166) over larger values. Configurations with $K \geq 3$ produced unstable clusters with fewer than 5% of samples.

5.2. Results

The two-cluster solution revealed a primary distinction along the productivity dimension (PC1) while showing similar editing behaviors (PC2) across groups:

High-volume writers (Cluster 1, 53.9%, n = 146): Higher keystroke activity, longer writing sessions, and more text production. Cluster center: PC1 = 2.22, PC2 = -0.33.

Low-volume writers (Cluster 2, 46.1%, n = 125): Lower keystroke counts and shorter writing sessions, but similar revision strategies to those of high-volume writers. Cluster center: PC1 = -2.60, PC2 = 0.39.

Figure 2 visualizes the cluster separation along PC1 (productivity) versus PC2 (editing behavior).

Figure 2: K-means clustering results (K = 2) projected onto the first two principal components. Clusters separate primarily along PC1 (productivity dimension).

5.3. Interpretation

The two-cluster solution identified writers along a productivity dimension, with both clusters showing similar editing strategies. The silhouette score (0.166) indicates these clusters represent a productivity continuum rather than discrete categories. High-volume and low-volume writers differ primarily in output quantity (PC1) while employing comparable revision approaches (PC2).

This finding differs from those of previous studies and is likely due to the time constraint in our study. Zhang et al. (2019) found four clusters in 35-minute essays, including groups that differed in effort and persistence but not in typing skill. Talebinamvar and Zarrabi (2022) identified five profiles distinguished by strategic pausing patterns (initial long pauses, varied subsequent pauses). Under our 15-minute constraint, such strategic and motivational differences likely collapsed into a single productivity dimension, as writers lacked time to employ varied approaches and strategic pausing behaviors.

6. Discussion and Conclusion

6.1. Summary of Findings

Our analysis of 259 argumentative essays with keystroke logging revealed how students at German high schools compose text under time pressure. Most students (96%) wrote linearly without recursion, and 88% completed only introductions and main parts, omitting conclusions. Writing was characterized by fluent bursts of text production with minimal extended pauses, suggesting a forward-moving composition strategy rather than extensive planning or revision.

Zone-specific analyses revealed systematic differences in writing behavior. The main part was written faster with lower revision rates and higher

efficiency compared to the introduction. This pattern suggests students use the introduction phase to orient themselves to the task and establish their argumentative stance, enabling more fluent writing as they develop their arguments.

An exploratory cluster analysis identified two writer profiles based on productivity: high-volume and low-volume writers who employed similar strategies but differed in output quantity. Exploratory correlational analyses revealed limited relationships between writing behaviors and academic grades, with only mathematics grades being significantly associated with conclusion completion, suggesting that writing behavior under time constraints may not strongly reflect broader academic performance.

6.2. Conclusion and Future Work

The GENIUS Keylog corpus offers a publicly available dataset that combines keystroke logging data with linguistic annotations of argumentative structure. To the best of our knowledge, this represents the first such resource. This combination makes it possible to investigate research questions that were previously difficult to address: How do writing processes relate to argumentative structure? Do revision patterns differ across essay sections? How do students manage time pressure during argumentative writing?

A methodological contribution is our approach for linking keystroke events to linguistic annotations through a fuzzy matching algorithm that enables temporal alignment between writing processes and annotated content zones.

The corpus provides baseline statistics and methodologies for studying timed argumentative writing, establishing a foundation for future investigations into how writing processes and textual structure interact.

We are currently collecting an expanded corpus with longer time limits and data from additional languages. This will allow us to investigate whether the patterns observed here—particularly the dominance of linear writing strategies—reflect the universal constraints of argumentative writing or are artifacts of time pressure and language-specific factors. The methodology established here can be applied to these new datasets to examine how writing processes vary across conditions.

Limitations

The design of this study, while producing valuable data on time-constrained writing, implies suggestions for future research:

Corpus size: The dataset comprises only 259 essays, which might not be sufficient for most machine learning purposes.

Time constraint effects: The 15-minute limit resulted in highly linear writing strategies and incomplete essays. While this captures authentic time-pressure behavior common in educational settings, it limits observations of recursive writing and complex revision strategies.

Task motivation: The voluntary nature of the task may have affected student engagement and effort. Higher-stakes assessments might elicit different writing behaviors.

Population homogeneity: The sample consists of German high school students who wrote argumentative essays on assigned topics; this limits its generalizability to other populations, genres, or contexts.

Ethics Statement

Our dataset contains essays written by underage students in the German school system as part of a 90-minute school lesson. We informed both the students and their parents about the collection of their written texts, as well as of any voluntarily given data such as age, gender, and grade, and we obtained their written consent to use and publish the data for research purposes.

7. Bibliographical References

R. Conijn, A. Rossetti, N. Vandermeulen, and L. Van Waes. 2025. Phase to phase: Developing an automated procedure to identify and visualize phases in writing sessions using keystroke data. *Journal of Writing Research*, 17(2):339–369.

Scott A. Crossley, Perpetual Baffour, Yu Tian, Aigner Picou, Meg Benner, and Ulrich Boser. 2022. The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (persuade) corpus 1.0. *Assessing Writing*, 54:100667.

Mariëlle Leijten, Luuk Van Waes, Iris Schrijver, Sarah Bernolet, and Lieve Vangehuchten. 2019. Mapping master's students' use of external sources in source-based writing in l1 and l2. *Studies in Second Language Acquisition*, 41(3):555–582.

Cerstin Mahlow. 2014. Learning from errors: Systematic analysis of complex writing errors for improving writing technology. In *Language Production, Cognition, and the Lexicon*, pages 419–438. Springer.

Cerstin Mahlow. 2015. Learning from errors: Systematic analysis of complex writing errors for improving writing technology. In N. Gala, R. Rapp, and G. Bel-Enguix, editors, *Language Production, Cognition, and the Lexicon*, volume 48 of *Text, Speech and Language Technology*, pages 433–457. Springer, Cham.

Tanja Riemeier, Claudia Aufschnaiter, Jan Fleischhauer, and Christian Rogge. 2012. Argumentationen von schülern prozessbasiert analysieren: Ansatz, vorgehen, befunde und implikationen. *Zeitschrift für Didaktik der Naturwissenschaften*, 18:141–180.

Nils-Jonathan Schaller, Yuning Ding, Thorben Jansen, and Andrea Horbach. 2025. Don't score too early! evaluating argument mining models on incomplete essays. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 345–355, Vienna, Austria. Association for Computational Linguistics.

Nils-Jonathan Schaller, Andrea Horbach, Lars Ingver Höft, Yuning Ding, Jan Luca Bahr, Jennifer Meyer, and Thorben Jansen. 2024. DARIUS: A comprehensive learner corpus for argument mining in German-language essays. In *Proceedings of Mahlowuistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4356–4367, Torino, Italia. ELRA and ICCL.

Ronja Schiller, Johanna Fleckenstein, Lars Höft, Andrea Horbach, and Jennifer Meyer. 2025. On the role of engagement in automated feedback effectiveness: Insights from keystroke logging. *Computers & Education*, 238:105386.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Manfred Stede, Sara Mamprin, Andreas Peldszus, André Herzog, David Kaupat, Christian Chiarcos, and Saskia Warzecha. 2015. *Handbuch Textannotation*. Potsdamer Kommentarkorpus 2.0. Universitätsverlag Potsdam.

M. Talebinamvar and F. Zarrabi. 2022. Clustering students' writing behaviors using keystroke logging: a learning analytic approach in efl writing. *Language Testing in Asia*, 12(6).

Y. Tian, S. Crossley, and L. Van Waes. 2025. The klicke corpus: Keystroke logging in compositions for knowledge evaluation. *Journal of Writing Research*, 17(1):23–60.

N. Vandermeulen, E. Van Steendam, S. De Maeyer, et al. 2024. Learning to write syntheses: the effect of process feedback and of observing models on performance and process behaviors. *Reading and Writing*, 37:1375–1405.

Georgios Velentzas, Andrew Caines, Rita Borgo, Erin Pacquetet, Clive Hamilton, Taylor Arnold, Diane Nicholls, Paula Buttery, Thomas Gaillat, Nicolas Ballier, and Helen Yannakoudakis. 2024. Logging keystrokes in writing by English learners. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10725–10746, Torino, Italia. ELRA and ICCL.

Åsa Wengelin. 2006. *Examining Pauses in Writing: Theory, Methods and Empirical Data*, volume 18, pages 107–130. Elsevier, United States. The information about affiliations in this record was updated in December 2015. The record was previously connected to the following departments: Linguistics and Phonetics (015010003).

M. Zhang, M. Zhu, P. Deane, and H. Guo. 2019. Identifying and comparing writing process patterns using keystroke logs. In M. Wiberg, S. Culpepper, R. Janssen, J. González, and D. Molenaar, editors, *Quantitative Psychology*, volume 265 of *Springer Proceedings in Mathematics & Statistics*, pages 459–475. Springer, Cham.

A. Appendix

Table 9: Grade Differences by Conclusion Completion

Subject	With Conclusion M (SD)	Without Conclusion M (SD)	n With	n Without	Difference	p
DEU	2.33 (0.96)	2.58 (0.97)	30	227	-0.24	.182
ENG	2.27 (0.87)	2.38 (0.88)	30	226	-0.12	.477
MAT	2.13 (0.97)	2.59 (1.08)	30	226	-0.46	.026*
CHE	2.29 (0.91)	2.56 (0.86)	24	200	-0.27	.153
PHY	2.23 (1.03)	2.37 (0.98)	26	201	-0.14	.493
BIO	2.17 (0.85)	2.38 (0.96)	29	217	-0.21	.211

Note: Lower grades indicate better performance in the German grading system. * $p < .05$

Table 10: Correlations Between Total Writing Output and Grades

Subject	r	p	Mean Output (chars)	SD
DEU	-.098	.125	1,847	982
ENG	.021	.737	1,845	978
MAT	-.116	.068	1,841	982
CHE	-.157	.021*	1,891	1,006
PHY	-.197	.003**	1,869	991
BIO	-.116	.073	1,860	989

Note: Negative correlations indicate more output associated with better grades. * $p < .05$; ** $p < .01$

Table 11: Correlations Between Writing Speed and Grades

Subject	r	p	Mean Speed (chars/sec)	SD
DEU	-.110	.084	1.54	0.87
ENG	.015	.818	1.54	0.87
MAT	-.086	.182	1.54	0.87
CHE	-.128	.062	1.55	0.88
PHY	-.160	.019*	1.54	0.87
BIO	-.101	.122	1.54	0.87

Note: Negative correlations indicate faster writing associated with better grades. Outliers beyond the 99th percentile were removed from analysis. * $p < .05$