FEAT-writing: An Interactive Training System for Argumentative Writing

Yuning Ding[1], Franziska Wehrhahn[1], Andrea Horbach[1,2,3]
yuning.ding@fernuni-hagen.de | f.wehrhahn@mail.de | horbach@leibniz-ipn.de
[1]CATALPA, FernUniversität in Hagen, Germany
[2]Leibniz Institute for Science and Mathematics Education, Kiel, Germany
[3]University of Kiel, Germany

Abstract

Recent developments in Natural Language Processing (NLP) for argument mining offer new opportunities to analyze the argumentative units (AUs) in student essays. These advancements can be leveraged to provide automatically generated feedback and exercises for students engaging in online argumentative essay writing practice. Writing standards for both native English speakers (L1) and English-as-a-foreign-language (L2) learners require students to understand formal essay structures and different AUs. To address this need, we developed FEAT-writing (Feedback and Exercises for Argumentative Training in writing), an interactive system that provides students with automatically generated exercises and distinct feedback on their argumentative writing. In a preliminary evaluation involving 346 students, we assessed the impact of six different automated feedback types on essay quality, with results showing general improvements in writing after receiving feedback from the system.

1 Introduction

Argumentative writing is a critical skill for academic success, requiring students to construct well-reasoned arguments, link ideas coherently, and support claims with relevant evidence. However, many students struggle to develop these skills (Graham and Perin, 2007). Mastering argumentative writing requires understanding essay structure and organizing ideas based on a clear argumentative framework (Hillocks, 2011).

One widely recognized model for teaching and analyzing argumentation is Toulmin's model (Toulmin, 2003), in which the central argumentative unit (AU) is a *claim* supported by *data* based on a *warrant*. This model has been incorporated into writing standards for L1 students, such as the Common Core State Standards for English Language Arts (CCSSO and NGA, 2010), which require students to introduce *claims* and logically organize the *evidence*. This model is also applied in L2 instruction, for instance, Germany's educational standards for the first foreign language (KMK, 2024) mandate that students express their own *opinions* and substantiate them with factual *reasons*.

Despite the importance of argumentative writing and the established standards, teachers in traditional classroom instruction often cannot provide detailed and individualized feedback due to time and resource constraints (Ferris, 2003). While automated tools exist to help students with grammar and spelling, they generally overlook the deeper aspects of language such as argumentation (Ranalli et al., 2017; Wilson and Roscoe, 2020).

To address these challenges, we developed FEAT-writing, an interactive system designed to help students improve their argumentative writing. The system generates exercises that aimed at supporting students to progressively build a solid argumentative framework, moving from simple tasks (i.e., distinguishing different AUs) to more complex ones (i.e., linking AUs with transitional words and supporting claims with evidence) and finally to writing complete argumentative essays. This approach is grounded in educational psychology, drawing from cognitive constructivism (Kalina and Powell, 2009) and a bottom-up learning approach (Sun et al., 2001). The system also provides students with automatically generated formative, summative, and elaborate feedback (Johnson and Priest, 2014), as well as automatically identified AUs visualized in color-coding (Maldonado-Otto and Ormsbee, 2019), which follows the principles of multimedia learning (Mayer, 2005).

We developed a web-based application (see Figure 1) that offers English writing exercises, aimed at helping students master the AUs and flow of argumentative writing. The final essay writing step in our system, writing an argumentative essay, was evaluated with 346 students in Germany,

Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations, pages 217–225
January 19–24, 2025.

and preliminary findings indicate improvements in the completeness of the argument structure after receiving color-coded elaborate feedback.

2 Background and Related Work

In this section, we first introduce the argumentative units types (AUs) used in our system, then review related work in two areas: feedback systems supporting the learning of AUs and exercises that train students to use them in writing.

2.1 AUs and their Effectiveness

The data foundation for our system is the PERSUADE corpus (Crossley et al., 2022, 2024), which contains over 280,000 discourse annotations for over 25,000 argumentative essays. Its annotation of AUs follows Toulmin's model. To focus on the most critical elements of argumentative writing and increase the system's accuracy, we use a simplified version of the original PERSUADE AUs following Ding et al. (2024). They defined AUs as followed:

- *Lead*: an introduction that begins with a statistic, a quotation, a description, or some other device to grab the reader's attention and point toward the thesis.
- *Position*: an opinion or conclusion on the main question.
- *Claim*: a claim that supports the position, refutes another claim or gives an opposing reason to the position.
- *Evidence*: ideas or examples that support claims, counterclaims, or rebuttals.
- *Concluding Statement*: a conclusion that restates the claims

The PERSUADE corpus also provides effectiveness scores for each AU. For the generation of exercises in the first three steps, we only use effective AUs, which are defined as shown in Appendix A.1.

2.2 AU Feedback Systems

With the popularity of online learning platforms, providing automated feedback becomes more and more critical to support teachers (Cavalcanti et al., 2021). With a focus on writing skills, numerous systems have been developed to provide students with automated feedback, primarily in the form of scores, since Page's seminal paper (Page, 1966). Comprehensive overviews of these systems can be found in the literature reviews by Ke and Ng (2019) and Beigman Klebanov and Madnani (2020). Specifically targeting argumentation feedback systems, Kuhn et al. (2017) provide a detailed summary.

This paper focuses on the systems that provide feedback on AUs. Wambsganss et al. (2021) introduced ArgueTutor, an adaptive dialog-based learning system that offers personalized feedback for argumentative texts by analyzing individual argumentative components. Bai and Stede (2022) provide feedback on the similarity between pairs of claims. The most similar work to ours is ALEN App (Wambsganss et al., 2022) and the system developed by Liu et al. (2016), both of which automatically detect claim-premise structures in students' essays and offer visual feedback to help students repair any broken argumentation structures.

However, these systems provide feedback only after students have completed a text or text snippets, whereas the training of argumentative writing benefits from step-by-step guidance. Our system, FEAT-writing, stands out by offering tailored task-specific feedback for exercises throughout the entire argumentative writing training process, addressing various stages of learning and providing the possibility of multiple attempts and revision.

2.3 AU Exercises and Automatic Generation

The automatic generation of language exercises is already a common application of NLP in education, encompassing vocabulary exercises (e.g. Heilman and Eskenazi, 2007; Peng et al., 2023), grammar exercises (e.g. Perez-Beltrachini et al., 2012; Heck et al., 2021) and their combination, such as c-tests (e.g. Haring et al., 2021; Lee et al., 2024). However, to the best of our knowledge, there has been little work on automating exercises specifically designed to train students in argumentative writing, or more specifically, in learning AUs.

Without automated generation, classroom exercises focusing on AUs, however, have demonstrated improvements in students' writing performance (Rafik-Galea et al., 2008; Khodabandeh et al., 2013). To address this gap, we leverage exercises such as distinguishing different AUs, linking AUs with transitional words, and supporting claims with the most effective evidence to enhance students' understanding and applications of AUs.

3 System Design

As introduced above and shown in Figure 1, FEAT-writing is designed to guide students through a

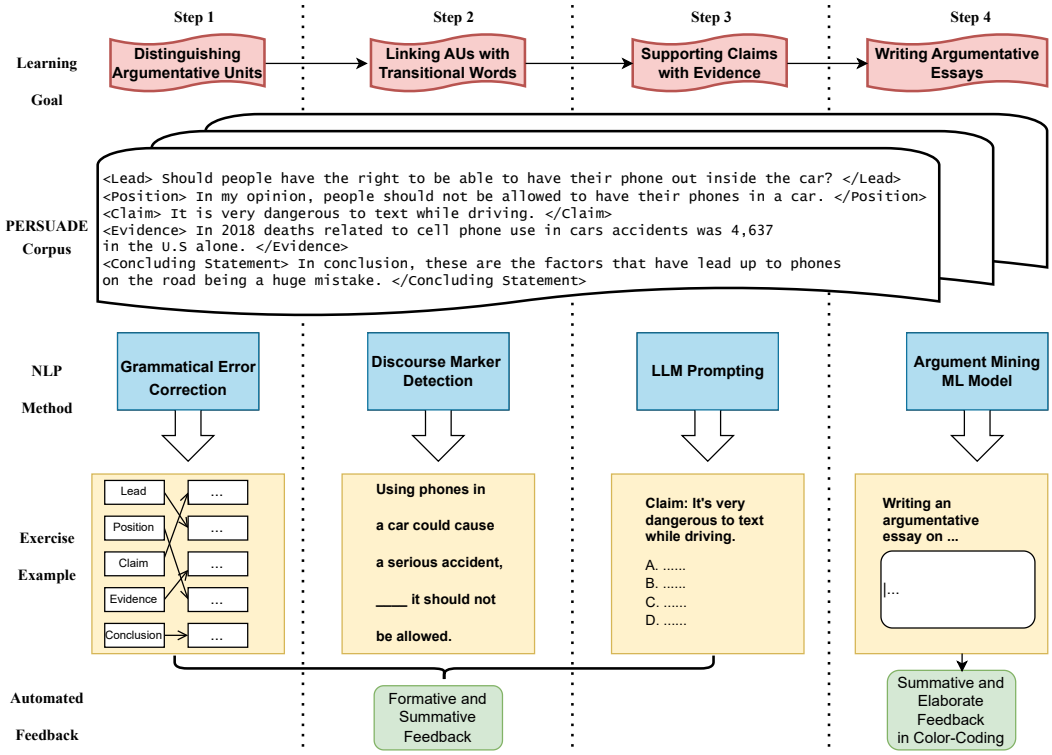

Figure 1: Structure of FEAT-writing.

sequence of exercises aimed at improving their argumentative writing skills. Each step builds upon the previous one, helping students move from basic recognition of AUs to writing full argumentative essays. This section outlines the system's functionality and the technology used in each step.

In alignment with scientific standards and the principles of Open Science, a key part of our plan involves the publication of anonymized datasets in publicly accessible repositories. To address privacy and ethical considerations, we have implemented a consent process. Before using the tool, students will be presented with a consent form allowing them to decide whether they permit the collection and publication of their data after anonymization.

A taskbar, displayed at the bottom of the page allows students to monitor their progress and switch between tasks. Students can easily return to prior exercises, practice further, and refine their skills, making the learning process more interactive and adaptive to individual needs.

3.1 Step 1: Distinguishing AUs

In the first step, students learn to distinguish between leads, positions, claims, evidence, and concluding statement by engaging in interactive linking tasks in a two-sided grid. The system presents definitions of AUs as described in Section 2.1 on the left side. On the right side, essay snippets constituting an AU, are selected from individual essays in the PERSUADE corpus so that all five types of AUs are represented and listed in random order. Using a click-and-link mechanism, students are asked to match each AU definition to its corresponding example. Once the task is completed and the "Check Answer" button is clicked, the system provides formative feedback by retaining the correct links and encouraging students to retry linking any previously incorrect ones. After three attempts with mistakes, the system prompts the student to view the correct answer. Upon finishing the current task correctly, students can either attempt another linking task or proceed to the next step.

To address spelling and grammatical errors in the raw texts from the PERSUADE corpus, we applied Grammatical Error Correction using LanguageTool[1]. This ensures that the example texts students work with are mostly free from distracting errors.

[1] https://languagetool.org

3.2 Step 2: Linking AUs with Transitional Words

In the second step, students practice linking AUs with appropriate transitional words to improve the logical flow of their arguments. Therefore, we filter the AUs used in step 1 for those containing discourse markers from a pre-compiled list. The system presents fill-in-the-blank tasks where students type in the correct transitional words for five different types of transitions: addition, contrast, cause and effect, example, and conclusion. The system allows alternatives with the same meaning.

After each attempt, the system provides formative feedback: following the first incorrect try, it reveals the transition type; after the second incorrect attempt, it suggests example transitional words to encourage further revision. If students enter an incorrect word three times, they are encouraged to click the "Show the Correct Answer" button, which provides summative feedback along with the correct solution.

3.3 Step 3: Supporting Claims with Evidence

In the third step, students focus on providing effective evidence to support claims. This is a multiple-choice exercise where the system presents a claim along with four potential pieces of evidence. Students need to select the evidence that best supports the claim, helping them understand the importance of using facts, statistics, and research to strengthen their arguments. After each selection, the system provides feedback on whether the chosen evidence is appropriate, explaining why certain pieces of evidence are not effective.

The system utilizes claim-evidence pairs from the AUs in Step 1. For the distractor choices, LLM prompting via the OpenAI API[2] is used to generate ineffective evidence. As the prompt shown in Appendix A.2, it also allows the system to provide immediate, detailed feedback on the effectiveness of the chosen evidence, teaching students to critically evaluate the support for their arguments.

3.4 Step 4: Writing Argumentative Essays

In the final step, students apply what they've learned in the previous exercises to write a full argumentative essay. After completing the essay, the system analyzes the text, identifying the AUs.

To achieve this, we utilize a machine learning pipeline for argument mining (Ding et al., 2022). In our process, 80% of the essays from the PERSUADE corpus, with annotated AUs, are preprocessed into tokens labeled with Inside-Outside-Beginning (IOB) tags and used as input for the pretrained Longformer model (longformer-large-4096) (Beltagy et al., 2020) for token classification. After 10 epochs of training with a maximum sequence length of 1024 tokens, the IOB-tagged tokens are transformed into predictions for different AUs during post-processing. The performance of this model is validated and tested on the remaining 10% of the essays, as discussed in Section 4.1.

Based on the AUs identified by this model in the students' essays, FEAT-writing provides two types of feedback as shown in Figure 2:

- **Summative Feedback** presented as a table listing the number of AUs identified in the text. This feedback offers a concise overview of the essay's structural completeness, giving students a clear, quantifiable measure of how well they have included key argumentative components (Tricomi and DePasque, 2016).
- **Elaborate Feedback** includes written text offering positive reinforcement, a list of AUs already present, and suggestions for additional AUs that students could incorporate in revision. The use of elaborate feedback is grounded in educational psychology, as providing detailed, constructive feedback can promote student self-efficacy and foster deeper learning. By offering specific guidance and motivation, this feedback type helps students understand not only what they are missing but also how to improve their writing (Cáceres et al., 2021).

Additionally, AUs are color-coded making it easier for students to understand the feedback and improve their writing (Maldonado-Otto and Ormsbee, 2019). After receiving feedback, students can revise their essays, return to a previous exercise, or write a new essay based on a different prompt.

4 Evaluation

FEAT-writing is evaluated primarily in the final step: writing argumentative essays. While the earlier exercises are crucial for building foundational skills in argumentative writing, their impact is best measured through the student's ability to produce

[2] https://platform.openai.com/docs/api-reference/introduction

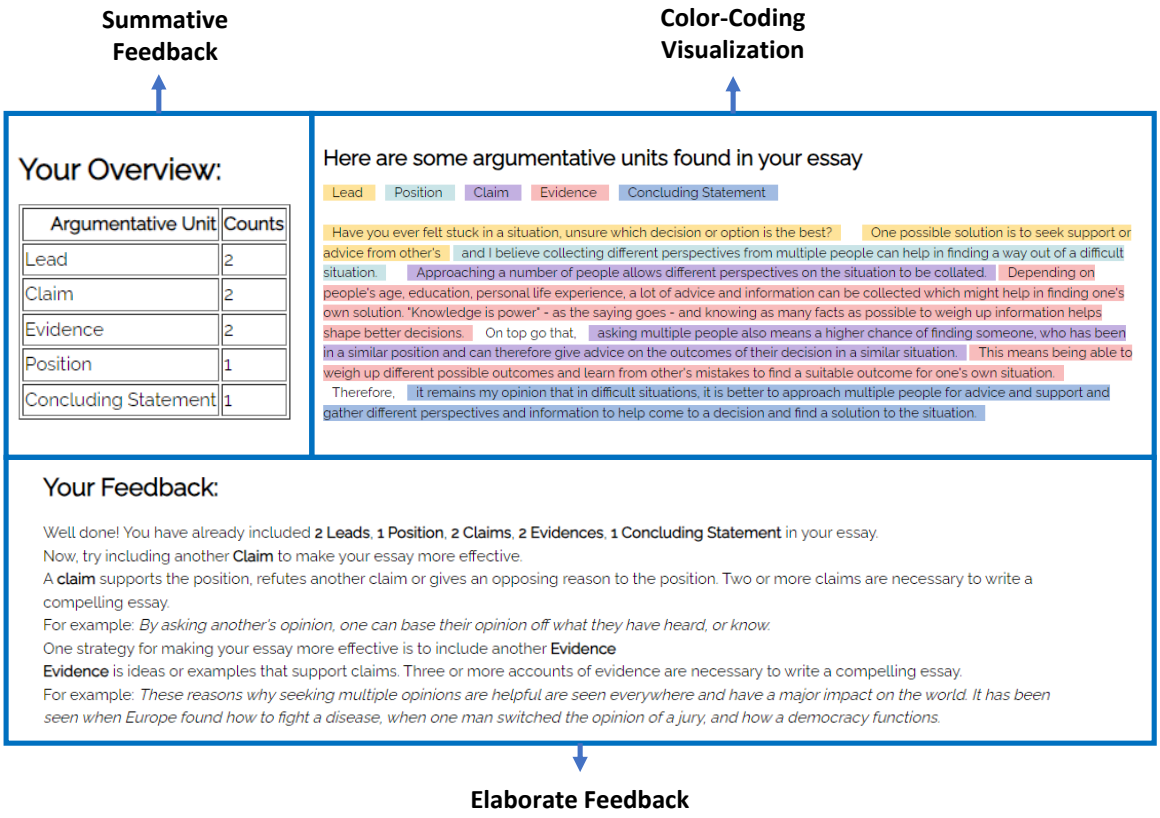

Figure 2: Feedback in Step 4 of FEAT-writing.

well-structured essays. Consequently, we first focus our evaluation on the system's performance in detecting AUs within student essays and the overall effectiveness of the feedback provided to students, leaving the evaluation of early steps in future work.

4.1 Performance of AU Detection

The performance evaluation of our argument mining model was conducted on 10% of the essays in the PERSUADE corpus. We consider a predicted AU to be a true positive if it overlaps with the corresponding ground truth AU by more than 50%. Conversely, predictions that do not match any ground truth units are marked as false positives, and ground truth units without a corresponding prediction are marked as false negatives. Using this approach, our model achieved an overall F1 score of 0.66, with a precision of 0.68 and a recall of 0.64. This indicates a reasonable level of accuracy in detecting AUs within the essays, which is crucial for providing meaningful feedback to students.

However, we acknowledge that assigning equal importance to all tokens in the matching process is a simplification. Methods that assign different weights to content and function words or incorporate token position, as described in (Schmidt et al., 2024), could further refine evaluation metrics and improve precision.

4.2 Effectiveness of Feedback

The evaluation of the effectiveness of automated feedback was structured as an online study with a focus on how different feedback types influence students' revisions and overall writing quality.

4.2.1 Variables and Methods

The independent variables were the use of color-coding in marking AUs within the student text and the feedback types provided by FEAT-writing. The feedback type included three variants:

- **Outcome Feedback**: A percentage score based on the presence of key AUs.
- **Summative Feedback**: A table listing the number of AUs identified, and
- **Elaborate Feedback**: Written feedback, including positive affirmations, a list of AUs already present, and suggestions for additional AUs to include.

Participants were randomly assigned to one of six feedback groups in a 2x3 between-subjects design, based on the presence or absence of color-

coding and the feedback type received. After writing an initial draft in response to a given prompt, participants received automated feedback according to their group assignment. They were then asked to revise their essays within a set timeframe (15 minutes).

Different from the description in Section 3.4, we introduced an **Outcome Feedback** score during the evaluation phase to provide a straightforward, quantitative measure of essay completeness. Specifically, students received a score of 100% if their essay contained at least one lead, one position, three claims, three pieces of evidence, and one concluding statement. This scoring method served as an initial, objective measure of the presence of key argumentative elements, which allowed us to compare essays systematically during the evaluation. Consequently, the scores participants received for their first draft (1st score) and their revised draft (2nd score) were used as our first dependent variables to measure Completeness Gain.

However, we recognize that this metric captures only the structural completeness of an essay in a rigid, predefined manner. While it was useful in the evaluation phase to gain insights into how revisions impacted essay structure, its limitations led to its exclusion from the final system, which focuses on more objective and detailed feedback.

Other dependent variables include:

- **Completeness Gain**: The difference between the 1st and 2nd scores.
- **Edit Distance**: The Levenshtein Distance (Levenshtein, 1966) between the original and revised essays, measuring the extent of changes made.
- **Lexical Diversity Gain**: The change in the Type-Token Ratio (TTR) between the original and revised essays, reflecting variations in word usage.

Data was collected through a combination of self-report surveys and log data from the writing tasks. Besides demographic information, participants were also surveyed about their experience with argumentative writing and automated feedback systems.

4.2.2 Results

Feedback is Effective A total of 346 L2 students from various German universities participated in this study. On average, participants were 31.2 years old and enrolled in their 6^{th} semester. This higher-than-usual average age for university students can be attributed to the fact that most participants were enrolled in remote university programs, which typically attract older students balancing studies with professional or personal responsibilities. Notably, 42.2% had no prior experience with argumentative essay writing, and 54.6% had not received automated feedback before this evaluation.

Dependent Variables	Average (M)	Standard Deviation (SD)
1st Score	76.7%	18.0%
2nd Score	81.5%	16.0%
Completeness Gain	4.8%	13.5%
Edit Distance	435.7	531.53
Lexical Diversity Gain	0.02	0.03

Table 1: Average of dependent variables showing essay improvement after feedback.

As shown in Table 1, the average **1st score** for the initial drafts was 76.7%. After receiving feedback and revising their essays, the average **2nd score** increased to 81.5%. The **Completeness Gain** between the drafts averaged 4.8%, with 28.03% participants improving their score by up to 71.8%, while 19.94% showed a decrease. The **Edit Distance** averaged 435.7, indicating a substantial degree of revisions. Only 17.3% of students made no edits after receiving feedback. Additionally, the average TTR improved from 0.53 in the initial draft to 0.55 in the revision, reflecting an improvement in **Lexical Diversity**.

Comparison of Feedback Groups Given that the dependent variables were not normally distributed, we used Kruskal-Wallis tests (Kruskal and Wallis, 1952) to compare six feedback groups. The results, presented in Appendix A.3, show significant differences between groups, but cannot specify between which groups these differences occur.

Therefore, Dunn's tests (Dunn, 1961) for pairwise comparisons were subsequently used to locate the specific differences between groups and revealed several key findings: For **Completeness Gain**, the group receiving outcome feedback with color-coding showed the highest improvement. In terms of **Edit Distance**, the longest changes were observed in the group receiving elaborate feedback with color-coding. This suggests that more elaborate feedback with visual cues encourages more extensive revisions. For **Lexical Diversity Gain**, the most diverse word usage was observed in the group that received elaborate feedback without color-coding.

5 Conclusion and Outlook

We introduced FEAT-writing, an interactive training system designed to enhance student's argumentative writing skills. It guides students through a series of exercises that progressively build their understanding, connecting, supporting, and writing of AUs. In each step, our system provides students with different types of automated feedback.

The results of our evaluation, which focused on the final step, where students wrote and revised full argumentative essays, indicate that FEAT-writing positively impacts students' argumentative writing.

Future work will first focus on extending our evaluation to the earlier steps in the system. These steps are crucial for skill building, but their effectiveness is not as easy to capture as the completeness of the final essay. Additionally, we plan to enhance the natural language processing capabilities of the system, including scoring the general quality of the essays automatically and refining the feedback mechanisms based on usability.

Limitations and Ethical Considerations

While FEAT-writing demonstrates the potential to support students' argumentative writing skills, there are some limitations to consider. First, the NLP models underlying FEAT-writing were trained on the PERSUADE corpus, which may carry inherent biases reflective of the data's sources. This could potentially affect the system's ability to provide equitable feedback across diverse linguistic and cultural backgrounds.

Furthermore, the predefined criteria for scoring, such as the specific requirements for "completeness", may not align perfectly with every educational context, potentially limiting the system's adaptability.

Given that FEAT-writing collects and processes students' written texts for evaluation, data privacy is a primary ethical consideration. In compliance with GDPR[3] and other data protection regulations, all user data, including essay submissions and interaction logs, are anonymized before analysis. The system only stores data necessary for educational purposes and does not retain personal information beyond what is required for feedback generation. Additionally, students' consent is obtained prior to participation, and they are fully informed about how their data will be used. Users have the right to withdraw their consent and request data deletion at any time, ensuring that their privacy and autonomy are respected throughout the writing process.

[3] https://gdpr-info.eu

Acknowledgements

This work was partially conducted at "CATALPA - Center of Advanced Technology for Assisted Learning and Predictive Analytics" of the FernUniversität in Hagen, Germany.

References

Xiaoyu Bai and Manfred Stede. 2022. Argument similarity assessment in german for intelligent tutoring: Crowdsourced dataset and first experiments. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2177–2187.

Beata Beigman Klebanov and Nitin Madnani. 2020. Automated evaluation of writing – 50 years and counting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7796–7810, Online. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Martín Cáceres, Miguel Nussbaum, Fernando González, and Vicente Gardulski. 2021. Is more detailed feedback better for problem-solving? *Interactive Learning Environments*, 29(7):1189–1210.

Anderson Pinheiro Cavalcanti, Arthur Barbosa, Ruan Carvalho, Fred Freitas, Yi-Shan Tsai, Dragan Gašević, and Rafael Ferreira Mello. 2021. Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 2:100027.

CCSSO and NGA. 2010. Common core state standards for english language arts 6-12. *Common core state standards for English language arts literacy in history/social studies, science, and technical subjects.*

S.A. Crossley, Y. Tian, P. Baffour, A. Franklin, M. Benner, and U. Boser. 2024. A large-scale corpus for assessing written argumentation: PERSUADE 2.0. *Assessing Writing*, 61:100865.

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.

Yuning Ding, Marie Bexte, and Andrea Horbach. 2022. Don't drop the topic-the role of the prompt in argument identification in student writing. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 124–133.

Yuning Ding, Omid Kashefi, Swapna Somasundaran, and Andrea Horbach. 2024. When argumentation meets cohesion: Enhancing automatic feedback in student writing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17513–17524.

Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64.

Dana R Ferris. 2003. Response to student writing: Implications for second language students. *Lawrence Earlbaum Associates*.

Steve Graham and Dolores Perin. 2007. Writing next-effective strategies to improve writing of adolescents in middle and high schools.

Christian Haring, Rene Lehmann, Andrea Horbach, and Torsten Zesch. 2021. C-test collector: A proficiency testing application to collect training data for C-tests. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 180–184, Online. Association for Computational Linguistics.

Tanja Heck, Detmar Meurers, and Stephen Bodnar. 2021. Automatic generation of form-based grammar exercises from authentic texts.

Michael Heilman and Maxine Eskenazi. 2007. Application of automatic thesaurus extraction for computer generation of vocabulary questions. In *SLaTE*, pages 65–68.

George Hillocks. 2011. Teaching argument writing, grades 6–12. *Portsmouth, NH: Heinemanri*.

Cheryl I Johnson and Heather A Priest. 2014. 19 the feedback principle in multimedia learning. *The Cambridge handbook of multimedia learning*, page 449.

Cody Kalina and KC Powell. 2009. Cognitive and social constructivism: Developing tools for an effective classroom. *Education*, 130(2):241–250.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *IJCAI*, volume 19, pages 6300–6308.

Farzaneh Khodabandeh, Manochehre Jafarigohar, Hassan Soleimani, and Fatemeh Hemmati. 2013. The impact of explicit, implicit, and no-formal genre-based instruction on argumentative essay writing. *Linguistics Journal*, 7(1).

KMK. 2024. Bildungsstandards für die erste Fremdsprache (Englisch/Französisch) für den Ersten Schulabschluss und den Mittleren Schulabschluss.

William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.

Deanna Kuhn, Laura Hemberger, and Valerie Khait. 2017. *Argue with me: Argument as a path to developing students' thinking and writing*. Routledge.

Ji-Ung Lee, Marc E Pfetsch, and Iryna Gurevych. 2024. Constrained c-test generation via mixed-integer programming. *arXiv preprint arXiv:2404.08821*.

V Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Proceedings of the Soviet physics doklady*.

Ming Liu, Yi Li, Weiwei Xu, and Li Liu. 2016. Automated essay feedback generation and its impact on revision. *IEEE Transactions on Learning Technologies*, 10(4):502–513.

C Maldonado-Otto and C Ormsbee. 2019. Color-coding affect on writing instruction for students with learning difficulties. In *ICERI2019 Proceedings*, pages 11421–11432. IATED.

RE Mayer. 2005. Cognitive theory of multimedia learning.

Ellis B Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Zhenhui Peng, Xingbo Wang, Qiushi Han, Junkai Zhu, Xiaojuan Ma, and Huamin Qu. 2023. Storyfier: Exploring vocabulary learning support with text generation models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–16.

Laura Perez-Beltrachini, Claire Gardent, and German Kruszewski. 2012. Generating grammar exercises. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 147–156.

S Rafik-Galea, Siti Zaidah Zainuddin, and PV Galea. 2008. Learning to think critically the toulmin way. In *13th Seminar International Conference on Thinking*.

Jim Ranalli, Stephanie Link, and Evgeny Chukharev-Hudilainen. 2017. Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology*, 37(1):8–25.

Federico M. Schmidt, Sebastian Gottifredi, and Alejandro J. García. 2024. Identifying arguments within a text: Categorizing errors and their impact in arguments' relation prediction. *International Journal of Approximate Reasoning*, 173:109267.

Ron Sun, Edward Merrill, and Todd Peterson. 2001. From implicit skills to explicit knowledge: A bottom-up model of skill learning. *Cognitive science*, 25(2):203–244.

Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.

Elizabeth Tricomi and Samantha DePasque. 2016. The role of feedback in learning and motivation. In *Recent developments in neuroscience research on human motivation*, pages 175–202. Emerald Group Publishing Limited.

Thiemo Wambsganss, Andrew Caines, and Paula Buttery. 2022. Alen app: Persuasive writing support to foster english language learning. *BEA 2022*, page 134.

Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. Arguetutor: An adaptive dialog-based learning system for argumentation skills. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–13.

Joshua Wilson and Rod D Roscoe. 2020. Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, 58(1):87–125.

A Appendix

A.1 Definitions of Effective AUs

The following definitions of effective AUs come from the scoring rubric of PERSUADE corpus, which can be found at https://github.com/scrosseye/persuade_corpus_2.0/blob/main/argumentation_effectiveness_rubric.pdf.

- *Effective Lead*: The lead grabs the reader's attention and strongly points toward the position.
- *Effective Position*: The position states a clear stance closely related to the topic.
- *Effective Claim*: The claim is closely relevant to the position and backs up the position with specific points or perspectives. The claim is valid and acceptable.
- *Effective Evidence*: The evidence is closely relevant to the claim they support and back up the claim objectively with concrete facts, examples, research, statistics, or studies. The reasons in the evidence support the claim and are sound and well substantiated.
- *Effective Concluding Statement*: The concluding summary effectively restates the claims using different wording. It may readdress the claims in light of the evidence provided.

A.2 LLM Prompt for Generation of Ineffective evidence

Given the claim: "$claim", generate three pieces of ineffective evidence, that are irrelevant to the claim, or provide only a few valid examples, making unsubstantiated assumptions. The evidence generated should be used as distractors for effective evidence: "$effective evidence", so they should have similar lengths but significant differences in content. For each ineffective piece of evidence, explain why it is not effective.

A.3 Results of Kruskal-Wallis Tests

Dependent Variables	**Kruskal-Wallis Value** χ^2**(5, 346)**
Completeness Gain	12.11*
Edit Distance	12.25*
Lexical Diversity Gain	11.22*

Table 2: Comparison of six feedback groups measured by Kruskal-Wallis tests. Results with * indicate significant values $p < .05$.