AL: An Adaptive Learning Support System for Argumentation Skills

Thiemo Wambsganss
University of St.Gallen
St.Gallen, Switzerland
thiemo.wambsganss@unisg.ch

Christina Niklaus
University of St.Gallen
St.Gallen, Switzerland
christina.niklaus@unisg.ch

Matthias Cetto
University of St.Gallen
St.Gallen, Switzerland
matthias.cetto@unisg.ch

Matthias Söllner
University of Kassel
Kassel, Germany
University of St.Gallen
St.Gallen, Switzerland
soellner@uni-kassel.de

Siegfried Handschuh
University of St.Gallen
St.Gallen, Switzerland
University of Passau
Passau, Germany
siegfried.handschuh@unisg.ch

Jan Marco Leimeister
University of St.Gallen
St.Gallen, Switzerland
University of Kassel
Kassel, Germany
janmarco.leimeister@unisg.ch

ABSTRACT

Recent advances in Natural Language Processing (NLP) bear the opportunity to analyze the argumentation quality of texts. This can be leveraged to provide students with individual and adaptive feedback in their personal learning journey. To test if individual feedback on students' argumentation will help them to write more convincing texts, we developed AL, an adaptive IT tool that provides students with feedback on the argumentation structure of a given text. We compared AL with 54 students to a proven argumentation support tool. We found students using AL wrote more convincing texts with better formal quality of argumentation compared to the ones using the traditional approach. The measured technology acceptance provided promising results to use this tool as a feedback application in different learning settings. The results suggest that learning applications based on NLP may have a beneficial use for developing better writing and reasoning for students in traditional learning settings.

Author Keywords

educational applications, pedagogical systems, argumentation learning, adaptive learning

CCS Concepts

•**Applied computing** → **Interactive learning environments;** •**Computing methodologies** → **Natural language processing;** •**Human-centered computing** → *Laboratory experiments;*

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376732

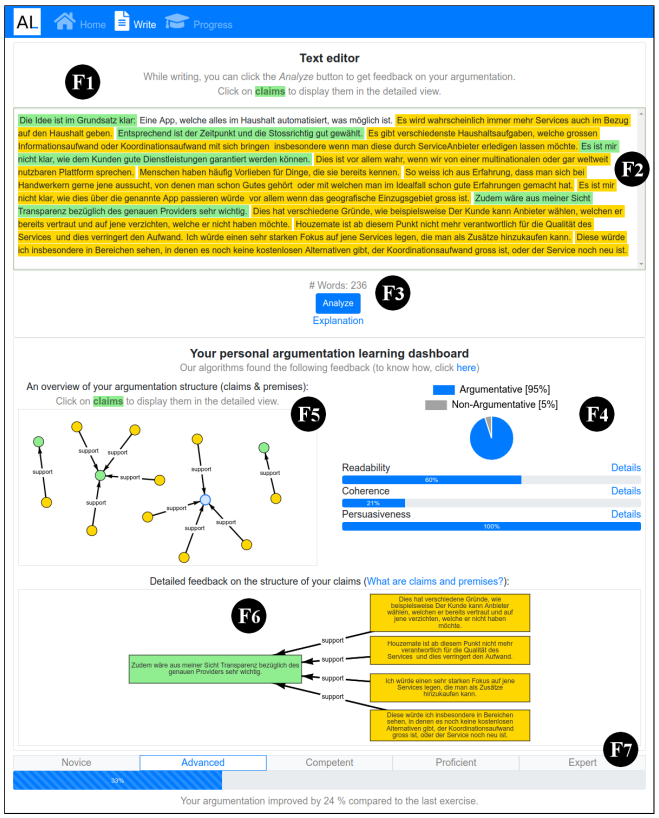

Figure 1. Screenshot of our adaptive learning support system: a user received feedback on the argumentation quality of her text

INTRODUCTION

Nowadays, information is readily available, so people need to develop skills other than the replication of information. This results in a shift of job profiles towards interdisciplinary, ambiguous and creative tasks [66]. Therefore, educational institutions need to evolve in their curricula, especially regarding the

compositions of skills and knowledge conveyed. In particular, teaching higher order thinking skills to students, such as critical thinking, collaboration or problem-solving, has become more important [17]. This has already been recognized by the Organization for Economic Co-operation and Development (OECD), which included these skills as a major element of their Learning Framework 2030 [41]. One subclass represents the skill of arguing in a structured, reflective and well-formed way [63]. Argumentation is not only an essential part of our daily communication and thinking but also contributes significantly to the competencies of communication, collaboration and problem-solving [34]. Starting with studies by Aristotle, the ability to form convincing arguments is recognized as the foundation for persuading an audience of novel ideas, and it plays a major role in strategic decision-making and analyzing different standpoints, especially with regard to managing digitally enabled organizations. To develop skills such as argumentation, it is of great importance for the individual student to receive continuous feedback throughout their learning journey, also called formative feedback [5, 27]. One of the major challenges is how to provide formative feedback in large-scale lectures effectively [23], since every student would need a personal tutor to have optimal learning conditions. However, this is naturally restricted by financial resources. One possible path for providing individual feedback is to leverage recent developments in Natural Language Processing and Machine Learning (ML). Researchers use Argumentation Mining (AM) to develop algorithms that extract argumentative components from given texts [37]. This information can be used to score the quality of a text and provide feedback concerning the persuasiveness of a text. Scientists, especially from the fields of educational technology, have designed tools to support the active teaching of argumentation for students with input masks or representational guidelines to enhance students' learning of argumentation (e.g., [10, 42, 45, 70]). However, current literature falls short of providing an approach with principles and proof on how to design an adaptive and intelligent IT tool to help students learn how to argue with intelligent formative feedback.

Given this potential for leveraging argumentation mining to enhance learning, we designed and built AL (short for *Argumentation Learning*), an adaptive learning tool that provides students with feedback on their argumentation structure during their writing process. We followed two different development approaches: a top-down approach, where we systematically analyzed literature in the field of educational technology and pedagogical theories based on [67] and interviewed 30 students with semi-structured interviews to rigorously derive requirements and principles for a first design of AL. Second, we followed a bottom-up approach, where we built low-fidelity prototypes of AL to test different design hypotheses with potential users to learn about the human interaction of an argumentation learning tool. With these two approaches, we present our final version of AL.

To design an individual and adaptive feedback tool, we collected a corpus of 1,000 student peer reviews from our lecture in which students give peers feedback on a digital business model. We wrote an annotation guideline and annotated the texts to build a corpus that fulfills our requirements. Afterwards, we trained a model to classify claims and premises and the discourse of those. This model now serves as the underlying feedback algorithm of AL. To determine the impact of AL on students' argumentation skills, we evaluated our learning tool in comparison with a carefully designed scripting tool, a proven approach for supporting argumentation in large-scale scenarios [19, 45]. In a study with 54 students, we observed that participants who used AL wrote formally more argumentative texts. Furthermore, the perceived persuasiveness of these texts was significantly higher than of the texts from the alternative tool. We also measured the technology acceptance of of both tools using key constructs [65, 64]. We found that the perceived usefulness and intention to use of AL provides promising results for its usage as a standard learning tool in lectures. The results suggest that AL helps students to write more structured texts and motivates them to write more persuasive texts in peer learning settings, such as peer feedback scenarios.

This work has three main contributions. First, AL is the first intelligent feedback learning tool for argumentation skills. Moreover, we show its effectiveness and usefulness through rigorously comparing AL with the current state of alternative learning tools for argumentation skills. The results demonstrate the benefits of leveraging NLP and ML for intelligent feedback on argumentation in a student's learning journey. Finally, our results show an exemplary case of supporting meta cognition skills in a scalable and individual way in possible large-scale scenarios. Thus, we provide design knowledge for other researchers and teachers to design and compare similar tools for supporting meta cognition skills of students.

RELATED WORK AND CONCEPTUAL BACKGROUND

Our work was inspired by previous studies on technology-mediated argumentation learning, by studies about argumentat mining algorithms and on cognitive dissonance theory, which serves as an underlying theory for our main hypothesis.

Technology-Mediated Argumentation Learning

Argumentation is an omnipresent foundation of our daily communication and thinking. In general, it aims at increasing or decreasing the acceptability of a controversial standpoint [15]. Logical, structured arguments are a required precondition for persuasive conversations, general decision-making and drawing acknowledged conclusions. As [34] states, the skill to argue is of great significance, not only for professional purposes like communication, collaboration and for solving difficult problems but also for most of our daily life. However, approaches for teaching argumentation are scarce. [29] identified three major causes for that: "*teachers lack the pedagogical skills to foster argumentation in the classroom, so there exists a lack of opportunities to practice argumentation; external pressures to cover material leaving no time for skill development; and deficient prior knowledge on the part of learners*". Therefore, many authors have claimed that fostering argumentation skills should be assigned a more central role in our formal educational system [13, 35]. Most students learn to argue in the course of their studies simply through

interactions with their classmates or teachers. In fact, individual support of argumentation learning is missing in most learning scenarios. However, to train skills such as argumentation, it is of great importance for the individual student to receive continuous feedback, also called formative feedback, throughout their learning journey [27]. According to [48], the outcome of feedback is a specific information relating to the task or process of learning that fills a gap between what is understood and what is aimed to be understood. Even in fields where argumentation is part of the curriculum, such as law and logic, a teacher's ability to provide feedback is naturally limited by constraints on time and availability. Especially in more common large-scale lectures, the ability to support a student's argumentation skills individually is hindered, since for teachers and professors, it is becoming increasingly difficult to provide ongoing and individual feedback to a single student [70]. The application of information technology in education bears several advantages, that is, consistency, scalability, perceived fairness, widespread use, better availability compared to human teachers, etc., and thus IT-based argumentation systems can help to relieve some of the burden on teachers to teach argumentation by supporting learners in creating, editing, interpreting or reviewing arguments [52]. This has been investigated across a variety of fields, including law [45], science [62, 42], and conversational argumentation [10]. Different technological approaches have been used in education. Especially intelligent tutoring systems (ITS) and computer-supported collaborative learning (CSCL) [31] are of special relevance for argumentation learning, since argumentative discussions and debates have been identified as a key for collaborative learning settings. Therefore, argumentation emerged as a focus area in CSCL. ITS is more centered around analyzing, modeling and supporting IT-based learning activities in specific domains. A relatively new research area is the combination of CSCL and ITS to support collaboration and argumentation in an adaptive and individual way [19]. Following [52], three different IT-based argumentation learning systems in the field of CSCL and ITS can be distinguished:

- **Representational guidance approaches** (e.g., [40, 45]) try to leverage argumentation learning by providing representations of argumentation structures with the objective to stimulate and improve individual reasoning, collaboration and learning. A common approach is to enable students to represent their argument structure in the form of node-and-link graphs.
- **Discussion scripting approaches** aim to provide structured elements for argumentation learning processes with the objective to foster interactions based on script theory of guidance [19]. A typical approach is to let students choose between predefined sentence openers when composing new text content [28].
- **Adaptive support approaches** (e.g., [45, 57, 59, 70]) describes a rather new field of argumentation support. The aim is to provide pedagogical feedback on a learner's actions and solutions, hints and recommendations to encourage and guide future activities in the writing processes or automated evaluation to indicate whether an argument is syntactically and semantically correct. However, as [51] describes, "*rigorous empirical research with respect to adaptation strategies is almost absent; a broad and solid theoretical underpinning, or theory of adaptation for collaborative and argumentative learning is still lacking*".

Our tool combines two approaches: adaptive feedback and representational guidance. We rely on NLP and ML to analyze the given text and provide adaptive feedback and an automated graph-based representation. We evaluate our tool against the discussion scripting approach, since it is most widespread and has been empirically proven to support students' formal quality of argumentation [19].

Argumentation Mining

An argument is a set of statements made up of three parts: a claim, a set of evidence or premises (e.g., facts) and an inference from the evidence to the claim [63]. Claim and premise represent the argument components. The claim is the central component of an argument, representing an arguable text unit, while the premises are propositions that either support or attack the claim, underpinning its plausibility. Support and attack are argumentative relations that model the discourse structure of arguments. Accordingly, an argument consists of one or more premises leading to exactly one conclusion, while argumentation connects together several arguments, thus establishing chains of reasoning, where claims are used as premises for deriving further claims. In that way, a regulated sequence of text with the goal of providing persuasive arguments for an intended conclusion or decision is constructed. AM is a research field in computational linguistics, gaining momentum in a lot of areas, including the legal domain [39], newswire articles [7, 11, 49], user-generated web content [69, 68, 26, 53, 30, 1], or online debates [6, 14]. AM aims at automatically identifying arguments in unstructured textual documents. Two main tasks can be distinguished: first, the detection of an argument, its boundaries and its relations with other text sections, and second, the detection and classification of the different components that make up the argument (i.e., the recognition of premises and conclusions). While the former requires a full argumentative text analysis in order to identify the global argumentation structure of the document at hand, the latter focuses on the internal structure of isolated arguments. In our approach, we will focus on the latter, carrying out the following subtasks:

- **Argument component classification**: classification of argumentative text into claims and premises
- **Argument relation classification**: identification of support relationships between pairs of argument components

Researchers have built increasing interest in intelligent writing assistance [55, 56, 57] since it enables argumentative writing support systems that provide tailored feedback about arguments in student essays. However, the complexity of using this technology in a certain teaching-learning scenario for educational purposes has rarely been assessed [59, 37].

Cognitive Dissonance

We built our research endeavor on cognitive dissonance theory. This theory supports our underlying hypothesis that individual and personal feedback on a student's argumentation motivates the student to improve her skill level. Cognitive dissonance

refers to the uncomfortable feeling that occurs when there is a conflict between one's existing knowledge or beliefs and contradicting presented information [18]. This unsatisfying internal state results in a high motivation to solve this inconsistency. According to Festinger's theory, an individual experiencing this dissonance has three possible ways to resolve it: change the behavior, change the belief or rationalize the behavior. Especially for students in a learning process, dissonance is a highly motivating factor to gain and acquire knowledge to actively resolve the dissonance [16]. It can be an initial trigger for a student's learning process and thus the construing of new knowledge structures [44] through critical reflection, also reflected in literature on transformation learning (e.g., [38]). However, the right portion of cognitive dissonance is very important for the motivation to solve it. According to Festinger, individuals might not be motivated enough to resolve it if the dissonance is too obvious, whereas a high level of dissonance might lead to frustration. Therefore, we believe that the right level of feedback on a student skill, such as argumentation skills, could lead to cognitive dissonance and thus to motivation to change the behavior, belief or knowledge to learn how to argue.

DESIGN OF ADAPTIVE LEARNING SYSTEM

In this section, we will explain how we designed and built our learning tool AL based on continued user feedback. The basic user interaction concept of AL is illustrated in figure 2. A user is writing or inserting a certain text and receives individual feedback on their argumentation. Therefore, AL consists of two main parts: A responsive and user-centered interface and the feedback algorithm in the back end.

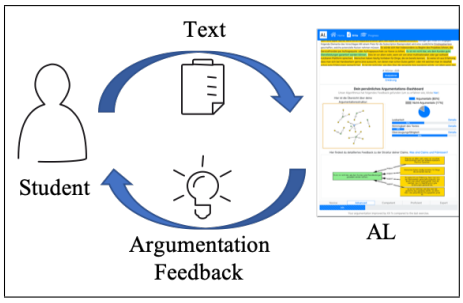

Figure 2. Basic user interaction concept of AL: students receive ongoing feedback on the argumentation of a given text

User Interface of AL

Deriving Requirements from Literature and Users

To build a user-centered learning tool, we followed two different approaches: a more rigorous top-down approach and an agile bottom-up approach following the build-measure-learn paradigm [46]. For the top-down approach, we first derived a set of meta-requirements from the current state of scientific literature for the design of an argumentation learning tool. To do so, a systematic literature search was conducted using the methodological approaches [9] and [67]. We initially focused our research on studies that demonstrate the successful implementation of learning tools for argumentation skills. Two broad areas for deriving requirements were identified: educational technology and pedagogical theories. Since the creation of a learning tool for argumentation skills is a complex project that is studied by psychologists, pedagogues and computer scientists with different methods, we first concentrated on these literature streams. We only included literature that deals with or contributes to a kind of learning tool in the field of argumentation learning, such as an established pedagogical theory. On this basis, we selected 67 papers for more intensive analysis. We have summarized similar topics of these contributions as literature issues and formed six clusters from them, which served as meta-requirements for IT learning tools for meta cognition skills. Next, we conducted thirty semi-structured interviews with students, using the expert interview method by [25]. The interview guideline consisted of 29 questions and each interview lasted around 30 to 50 minutes. The interviewees were a random subset of the population of students at our university who are all potential users of an argumentation learning tool. The participants were asked about the following topics: experience with technology-based learning systems, perception of existing learning systems in use, importance of skills in university education, requirements for a system that supports learning meta cognition skills (e.g., functionalities, design) and requirements for a system that supports learning how to argue (e.g., functionalities, design). The interviewed students were between 22 and 28 years old and all students of economics, computer science or psychology. 13 were male, 17 female. After a more precise transcription, the interviews were evaluated using a qualitative content analysis. The interviews were coded, and abstract categories were formed. The coding was performed using open coding to form a uniform coding system during evaluation [25]. Based on these results, we gathered 180 user stories, aggregated the most common ones and identified ten user requirements following [8]. From those user requirements and the derived meta-requirements from our systematic literature review, we concluded several design principles that influenced the design of AL.

Besides the rigorous approach, we followed a continued bottom-up approach at the same time. We built low-fidelity prototypes of AL to test different design hypotheses with potential users to learn about the human interaction of an argumentation learning tool. We started the testing with three low-fidelity paper prototypes and later with two digital mock-ups of AL. For example, we hypothesized that users aim to receive visual feedback of their argumentation. We tested this with a paper prototype that provided visual feedback on the argumentation of a given text (simulating the feedback algorithm by a human). The hypothesis was validated with 10 users, which overall liked the concept of visual feedback. Therefore, the final prototype of AL now contributes to that with a graph engine that provides a visual representation of the argumentation of an analyzed text. In total, we conducted five cycles with a total of 49 different users (around ten users per cycle). These users were different to the ones recruited for the semi-structured interviews but also students from our university with a similar age and gender distribution. Based on those two approaches, we finally came up with seven design principles on how to build an adaptive argumentation feedback tool illustrated in table 1. The design principles were instantiated with our current version of AL.

	Design Principle
1)	Provide the learning tool with a learning progress indicator in order for users to actively monitor their past and current learning development to convey a goal and purpose of learning for a long-term learning.
2)	Provide the learning tool as a web-based application with a responsive, lean and intuitive UX in order for users to intuitively and enjoyably use the tool.
3)	Provide the learning tool with a learning dashboard using gamification elements and a choice of different granularity levels in order for users to receive the right amount of needed feedback information.
4)	Provide the learning tool with a function that displays the theory of argumentation before arguing and incorporate it into the feedback in order to have an orientation in learning.
5)	Provide the learning tool with visual argumentation and discourse feedback on written or spoken information in order for users to apply argumentation and receive instant and individual feedback at any time and any place.
6)	Provide the learning tool with argumentation feedback along best practices, examples based on theory and/or *how-to-argue* guidelines and do not compare argumentation.
7)	Provide the learning tool with adaptive and individual feedback in order for users to receive useful and specific feedback on their given argumentation.

Table 1. Design principles on how to build an adaptive argumentation feedback tool

User Interaction of AL

Following above mentioned design principles, AL is built as a responsive web-based application that can be used on all kind of devices. A screenshot of AL and its different functionalities (e.g., *F1* - *F7*) can be seen in figure 1. AL provides the user with a rather simple text input field (*F1*) with a word count (*F3*) in which they can write or copy a text. Below the input field, the user receives feedback on the argumentation structure of their text in a personal learning dashboard (*F4* and *F5*). The dashboard provides different granularity levels of feedback, which enables the user to control the amount of needed feedback information [50]. A visual graph-based representation of the argumentation structure of a given text (*F5*) and three summarizing scores give an initial overview of the quality of the text (*F4*). In the written text, the identified claims are colored in green and the premises are highlighted in yellow to provide the user with an instant feedback on their own given input (*F2*). By clicking on the marked text fields or on the nodes in the graph, a more detailed view of the discourse of the argument will appear (*F6*). This shows clearly if a claim is sufficiently supported (as in *F6*) or if it misses a premise (see in figure 3 top). This function provides the user with clear action steps on how to improve the persuasiveness and formal quality of their texts. Moreover, best practices and explanations about argumentation and argumentation theory are provided by clicking on the "*Explanation*" or "*Help*" button (see figure 3). The three summarizing scores readability, coherence and persuasiveness (*F4*) provide the student with a ranking of their text to provide superficial instant feedback. By clicking on the scores or on "*details*", the methodology for calculating the scores, as well as concrete hints, action steps and explanations on how the student can increase her score level will be shown (see figure 3 below). These action steps provide the user with orientation and context to improve their writing quality [54, 27]. On the bottom of the tool, AL provides a learning progress bar (*F7*), which actively monitors the student's past and current learning development to convey a goal and purpose of learning [54, 27]. Based on our user studies, AL is not provided with a social comparison of the user's argumentation level with peers, since we often received negative feedback for such a functionality. In fact, the users wanted to receive as individual of a feedback as possible based on theory, since the level of argumentation is very context-sensitive. Therefore, we provided AL with an intelligent feedback algorithm that provides adaptive and individual feedback.

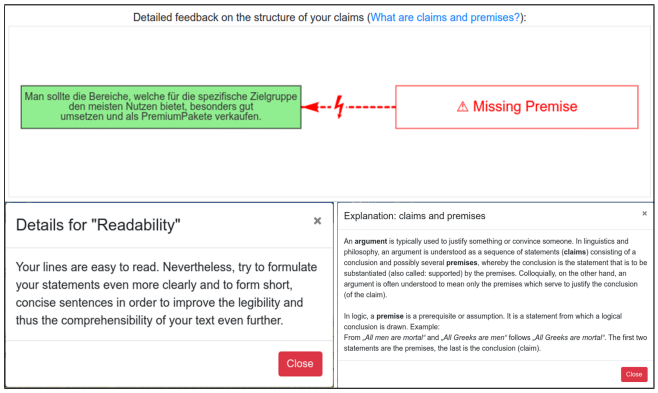

Figure 3. Top: screenshot of detailed discourse feedback of unsupported claim. Below: screenshot of exemplary explanations and details of AL

Feedback Algorithm of AL

To design an individual and adaptive feedback tool, we collected and annotated our own corpus to train and tune a model that fulfils the users' requirements to give instant feedback on their texts.

Building a Corpus of Persuasive Student Essays

A major prerequisite for developing NLP methods that are able to identify argument components and argumentative relations in written texts is the availability of annotated corpora. Since no suitable corpus was available that A) contained annotated persuasive student essays in German, B) consisted of a sufficient corpus size to be able to use the trained model in a real-world scenario that fulfils our user requirements and C) followed a novel annotation guideline for guiding the annotators towards an adequate agreement, we decided to build our own data set. Therefore, we collected a corpus of 1,000 student-generated peer reviews written in German. The data was collected in one of our mandatory business innovation lectures in a master program at our university. In this lecture, around 200 students develop and present a new business model for which they receive three peer feedback reviews each where a student from the same course elaborates on the strengths and weaknesses of a business model and gives persuasive recommendations on what could be improved. We collected a random subset of 1,000 of these reviews from around 7,000 documents form the last years. In the annotation process, we followed the approach described in [58]. Three native German speakers were hired to annotate the reviews independently from each other for *claims* and *premises* as well as their *argumentative relationship* in terms of support

and attack, according to the annotation guidelines we specified. Inspired by [58], our guidelines consisted of 15 pages, including definitions and rules for what is an argument, which annotation scheme is to be used and how argument components and argumentative relations are to be judged. Several training sessions were performed to resolve disagreements among the annotators and to reach a common understanding of the annotation guidelines. We used the brat rapid annotation tool, since it provides a graphical interface for marking up text units and linking their relations [61]. After the first 100 reviews were annotated by all three annotators, we calculated the inter-annotator agreement (IAA) scores (see table 2). As we obtained satisfying results, we proceeded with a single annotator who marked up the remaining 900 documents.

To evaluate the reliability of the argument component and argumentative relations annotations, we followed the approach of [56]. With regard to the argument structure, two strategies were used. Since there were no predefined markables, annotators not only had to identify the type of argument component but also its boundaries. In order to assess the latter, we used Krippendorff's α_U [33], which allows for assessing the reliability of an annotated corpus, considering the differences in the markable boundaries. Moreover, we calculated percentage agreement and two chance-corrected measures, multi-π [20] and Krippendorff's α [32], to evaluate the annotators' agreement in terms of the selected category of an argument component for a given sentence. We decided to operate at sentence level, since only 4.9% of the sentences in the corpus contain annotations of different argument components, e.g., both a claim and a premise span. Thus, evaluating the reliability at sentence level served as a good approximation of the IAA. At the token level, the following class distribution was achieved: 42.8% claim, 45.0% premise and 12.2% are not annotated. At the level of individual sentences, 40.3% contain a claim, 40.6% a premise and 24.1% none-annotation. Hence, 5.0% of the sentences contain several annotations.

	%	Multi-π	Krippendorff's α	Krippendorff's α_U
Claim	0.7053	0.3423	0.3424	0.4379
Premise	0.7048	0.3738	0.3739	0.3812

Table 2. IAA of argument component annotations.

Table 2 displays the resulting inter-rater agreement scores. We obtained an IAA of 70.5% for both the claims and the premises. The corresponding multi-π as well as Krippendorff's α scores are 34.2% and 37.4%, indicating a fair agreement for both categories. The unitized α of the claim is considerably higher compared to the sentence-level agreement. The unitized α of the premise annotations is higher too, but only slightly. Thus, the boundaries of both claims and premises are more precisely identified in comparison to the classification into argument types. The joint unitized measure for both categories is $\alpha_U = 0.4096$, suggesting a moderate agreement between the annotators. Hence, we conclude that the annotation of the argument components in student-generated peer reviews is reliably possible. Since the number of attack relations is so small, we decided to focus on the support relations, distinguishing only between the two types support and non-support. To evaluate the reliability of argumentative relations, we used the set of all relations that were possible during the annotation task, i.e., all pairs between a claim and a premise and between two premises. We obtain an IAA score of 78.0% for the support relations, concluding that argumentative support relations can be reliably annotated in our corpus.

NLP and AM Pipeline

To provide students with feedback on the argumentation quality of their texts, we first of all implemented an approach for detecting arguments in them. This approach consists of two subtasks. In a first step, we identified the components of arguments in terms of claims and premises. Next, we determined whether there is an argumentative relation between a pair of argument components. To do so, we followed the approach described in [57], a state-of-the-art approach for identifying argumentative discourse structures in persuasive essays.

Subtask 1: Argument Component Identification

The identification of argument components is considered as a sentence-level multi-class classification task, where each sentence in the dataset is labeled as either *claim*, *premise* or *non-argumentative*. Hence, apart from classifying argument components as claims or premises, this task includes the separation of argumentative from non-argumentative text units. To ensure an equal distribution of classes among training and test sets in our experiments, we performed a stratified split of the data set into a 80% training set and a 20% test set, resulting in the distribution of 32% claims, 32% premises and 36% non-argumentative spans (for both training and test set). In accordance with [3], we used several classifiers (Support Vector Machine (SVM), Logistic Regression, Random Forest, Multinomial Naive Bayes (NB), Gaussian NB, Nearest Neighbor and AdaBoosted Decision Tree) for the task of argument component identification. To tune the parameters of our models, we applied grid search. For pre-processing the documents, we used the spacy parser.[1] The features given in table 3 were extracted for training a model to perform the task of argument component identification following the taxonomy of [24]. We found that an SVM achieves the best results, with an accuracy of 65.4% on the test set. The resulting argument structure is visualized directly in the student-generated text by highlighting claim components in green and premise components in yellow, while non-argumentative text spans are not marked up.

Group	Feature
Lexical	Unigrams Dependency Tuple
Structural	Token statistics Component position
Indicators	Type indicators First-person indicators
Contextual	Type indicators in context Shared phrases
Syntactic	Subclauses Depth of parse tree Tense of main verb Modal verbs POS distribution
Probability	Type probability
Discourse	Discourse Triples
Embedding	Combined word embeddings

Table 3. Features used for argument component identification [58]

[1] https://spacy.io/

Subtask 2: Argument Relation Identification
The identification of argumentative relations is considered a binary classification task of argument component pairs, where each pair is classified as either *support* or *non-support*. All possible combinations of argument components are tested. Like before, we randomly split the dataset in a 80% training set and a 20% test set and determined the best performing system using 10-fold cross-validation on the training set. We used the same pre-processing pipeline as described in the previous paragraph and extracted the features detailed in table 4. The comparison of several classifiers (SVM, Logistic Regression, Random Forest, Multinomial NB, Gaussian NB, Nearest Neighbor and AdaBoosted Decision Tree) revealed that an SVM achieves the best results for our corpus, obtaining an accuracy of 72.1% on the test set. To tune the parameters of our model, we again used grid search. The resulting argumentative discourse structures are visualized in terms of a directed graph, connecting a claim with its supporting premises (see figure 1, F5). Unsupported claims, i.e., claims that lack supporting evidence, are highlighted to point out that further support needs to be provided here (see figure 3).

Group	Feature
Lexical	Unigrams
Syntactic	Part-of-speech Production rules
Structural	Token statistics Component statistics Position features
Indicator	Indicator source/target Indicators between Indicators context
Discourse	Discourse Triples
PMI	Pointwise mutual information
ShNo	Shared nouns

Table 4. Features used for argument relation identification [58]

Besides, we calculated a number of summary scores for providing students with an overview of the quality of their argumentation based on previously extracted argumentative discourse structures, including

- **Readability**: How readable is the text based on the Flesch Reading Ease score [22]?
- **Coherence**: How large is the proportion of sentences that are connected via discourse markers?
- **Persuasiveness**: How large is the proportion of claims that are supported by premises as compared to unsupported claims?

Alternative Learning System: Discussion Scripting

To evaluate AL, we compared it to a discussion scripting application, which we also built ourselves. Implementing our own discussion scripting approach allowed us to control the differences and similarities in the design between the discussion scripting tool and AL. For the design we followed the approach of [60], since it is well-cited and empirically proven to foster the formal quality of argumentation of students. The learning tool supports the writing process of users with input masks (see figure 4 *F1* and *F2*). Users can use these input masks to compose a formally correct argument. By clicking "*Add*", the argument is attached to the text field (*F3*). If the user decides to not use the input masks, they can also write directly into the final text field. To keep AL and the discussion scripting approach consistent with each other, there are many functions that are shared between them. First, the introduction text is the same across both apps. The help and explanation buttons in the discussion scripting approach correspond respectively to the same buttons in AL. Moreover, both text fields consist of a word count function to provide guidance in the writing process. We used the design of the particular input mask of this scripting approach and tested it in our non-collaborative experimental use case to measure the influence of the approach on the argumentation quality of students' texts. However, both approaches, our adaptive system and the scripting approach, could be used in a collaborative learning scenario, e.g., where students give each other feedback on a business model and discuss these.

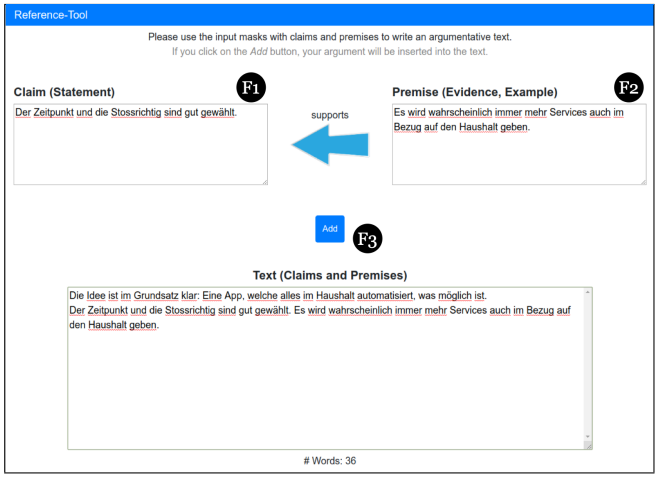

Figure 4. Screenshot of discussion scripting approach: a user enters a claim and the corresponding premise into the input masks and adds it to the final text

EXPERIMENTAL SETUP

In this section, we describe the experimental setup for our study. Its goal was to evaluate our hypothesis that individual feedback on student's argumentation will help them to write more convincing texts. To achieve our goal, we designed a laboratory experiment in which participants were asked to write a peer feedback based on a given essay. Participants were randomly assigned to treatment and control group. The treatment group used AL, while participants in the control group used the discussion scripting application.[2] We recruited 54 students from our university through social networks and mailing lists to take part in our experiment. After randomization, we happened to have 24 participants in the treatment and 30 in the control group. We invited them to the laboratory of our university, where we conducted the study on the exact same devices. Participants of the treatment group had an average age of 23.8 (SD= 3,86), 15 were male, 9 female. In the control group, participants' average age was 23.03 (SD= 2.12), 22 were male, 8 female. All participants were compensated with 15 USD for a 30 to 40 minutes experiment.

[2]AL was designed in German to provide German students with feedback on German texts. However, for ease of understanding in this paper, we translated our user interface into English (e.g., see figure 1).

Design and Procedure

The experiment consisted of three main phases: 1) pre-test phase, 2) individual writing phase and 3) post-test phase. The pre- and post-phases were consistent for all participants. In the writing phase, the treatment group used AL and the control group wrote a text using the alternative tool.

1) Pre-test phase: The experiment started with a pre-survey with 14 questions. Here, we tested three different constructs to assess whether the randomization resulted in randomized groups. First, we asked four items to test the personal innovativeness in the domain of information technology of the participants following [2]. Second, we tested the construct of feedback seeking of individuals following [4]. Example items are: "*It is important for me to receive feedback on my performance.*" or "*I find feedback on my performance useful.*" Both constructs were measured with a 1- to 5-point Likert scale (1: totally agree to 5: totally disagree, with 3 being a neutral statement). Third, we captured the construct of passive argumentative competency following the design of [21], since it is a proven construct to measure argumentative competencies in German. We wanted to control for the argumentative competencies, since we later measured the formal and perceived quality of argumentation of the written texts. Participants were asked to read a discussion of two teachers concerning the topic "*Does TV make students aggressive?*" We retrieved the topic with the discussion as well as the measurements from [21]. Based on the discussion, we asked the participants three questions concerning the argumentation structure and the content of the text with multiple choice answers: "*What kind of argumentation style or structure is used*?", "*How can a new argument be added to the discussion?*" and "*Which of the following standpoints do both parties agree on*?" [21]. Additionally, participants were asked how sure they were about the answers on a 1- to 5-point Likert scale (1: very sure, 5: not very sure, with 3 being a neutral statement). The competencies were then measured with a certain score from 0 to 27 following the measurements of [21].

2) Writing phase: In the writing phase of the experiments, we asked the participants to write a review about the argumentation of both parties (pro and contra) concerning the weaknesses and strengths of their argumentation. The participants were told to spend at least 15 minutes on writing this review. A countdown indicated them the remaining time. They were only able to continue the experiment after the countdown was finished. The treatment group was using AL to write the review, the control group was using the reference tool. We did not provide any introduction to any of the tools. The students using AL retrieved individual and adaptive feedback based on our feedback algorithms. Participants using the reference tool retrieved help based on input formats during the writing process.

3) Post-test phase: In the post-survey, we measured perceived usefulness, intention to use and ease of use following the technology acceptance model of [65, 64] and captured the demographics. In total, we asked 16 questions. Example items for the three constructs are: "*Imagine the tool would be available in your next course, would you use it?*", "*The use of the argumentation tool enables me to write better argumentative texts.*" or "*I would find the tool to be flexible to interact with*". Moreover, we were asking three qualitative questions: "*What did you particularly like about the use of the argumentation tool?*", "*What else could be improved?*" and "*Do you have any other ideas?*"

Measurement of Argumentation Quality

Besides measuring the technology acceptance, our main objective was to measure the quality of the written texts from both groups to evaluate our main hypothesis. Therefore, we measured two main variables: 1) the formal quality of argumentation and 2) the perceived quality of argumentation.

1) Formal quality of argumentation: The written peer reviews were analyzed for the formal quality of argumentation. We applied the annotation scheme for argumentative knowledge construction described by [71]. This annotation scheme was applied in various studies and has proven high objectivity, reliability and validity (e.g., [60]). To measure the formal quality of argumentation, the annotator had to distinguish between a) *unsupported claims*, b) *supported claims*, c) *limited claims,* and d) *supported and limited claims*. A more precise description of the scheme can be found in [71]. One annotator, who had already participated in the annotation process for our corpus, then annotated the received text from the participants based on our annotation guidelines and the experience before. We only relied on one annotator since an annotator agreement was already conducted during the corpus collection process for the same kind of texts in the same domain (peer feedbacks). The formal quality of argumentation of the individual user was then defined by the number of arguments written by a user during the writing phase. Following [60], only *supported*, *limited* and *supported and limited claims* were counted as argumentation.

2) Perceived quality of argumentation: The perceived quality of argumentation was annotated by two different annotators. The objective was to subjectively judge how persuasive the given argumentation is on a Likert scale from 1 to 5 points (1: very persuasive, 5: not very persuasive). Since this is a very subjective measurement, we took the mean of both annotators as a final variable for the perceived quality of argumentation of the texts.

EVALUATION AND RESULTS

To evaluate our hypothesis that individual feedback on students' argumentation will help them to write more convincing texts, we aim to answer two research questions (RQ):

RQ1: *Do students perceive AL to be useful and easy to use, and would they continue to use it in the future?*

RQ2: *How effective is AL with helping users to write more persuasive texts compared to the traditional, proven approach?*

The first research question will be answered by comparing the constructs of perceived usefulness, intention to use and ease of use for participants using AL compared to participants using the alternative tool. In particular, we will use a double-sided t-test to evaluate whether the means of the constructs are significantly different. Moreover, we will compare the

results of AL to the midpoints scale to validate a general positive technology acceptance as done in [36]. To evaluate the second research question, we compare the formal quality of argumentation as well as the perceived quality of argumentation between the written text of the treatment and the control group. We perform a double-sided t-test to assess whether differences between both groups are statistically significant. In order to control for potential effects of interfering variables with our rather small sample size and to ensure that randomization was successful, we compared the differences in the means of the three constructs included in the pre-test. For all three constructs, including personal innovativeness, feedback seeking of individuals and passive argumentative competency, we received p-values larger than 0.05 between the treatment and the control group. The p-value for personal innovativeness between both groups was p= 0.801, for feedback seeking of individuals p= 0.624, and for passive argumentative competency p= 0.375. This shows that no significant difference in the mean values for these three constructs exists between the groups.

Group	Intention to use	Perceived usefulness	Perceived ease of use
Mean AL	2.33	2.52	2.17
Mean reference tool	3.5	3.28	2.84
SD AL	0.59	0.58	0.65
SD reference tool	1.14	1.12	1.08
p-value	<0.001	0.006	0.012

Table 5. Results of the technology acceptance of AL and the reference tool on a 1 - 5 Likert Scale (1: high, 5: low)

Technology Acceptance

For the technology acceptance, we calculated the average of every construct. The answers were provided on a 1- to 5-point Likert scale (1: very sure, 5: not very sure). First, we compared the results of AL with the results of the alternative tool. The perceived usefulness of AL was rated with a mean value of 2.52 (SD= 0.58) and the average of perceived use of ease of AL was 2.17 (SD= 0.65). The mean value of intention to use of participants using AL as a writing tool was 2.33 (SD= 0.58). These values are significantly better than the results of the alternative scripting approach. For perceived usefulness we observed a mean value of 3.2 (SD= 1.12) and for perceived ease of use the value was 2.83 (SD= 1.08) for participants from the control group. The mean value for the intention to use was 3.5 (SD= 1.13). The results clearly show that the participants of our experiment rated the acceptance of AL as an adaptive feedback tool positively compared to the usage of the alternative application. The statistical significance was also proven in a double-sided t-test for all three constructs (see table 5). Moreover, the mean values of AL are also very promising when comparing the results to the midpoints. All results are better than the neutral value of 3. Especially the perceived usefulness for writing better argumentative texts and the intention to use AL as a writing support tool show promising results. A positive technology acceptance is especially important for learning tools to ensure students are perceiving the usage of the tool as helpful, useful and easy to interact. This will foster motivation and engagement to use the learning application. The perceived usefulness and intention to use provides promising results to use this tool as a feedback application in different learning settings.

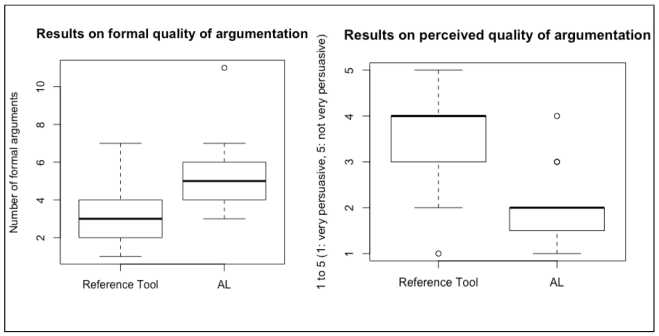

Figure 5. Results on formal (left) and perceived (right) quality of argumentation between both tools

Argumentation Quality of Written Texts

The mean number of arguments in texts from participants using AL was 5.08 (SD= 1.76). From the text of participants using the alternative tool, we counted a mean of 3.2 arguments (SD= 1.51) (see figure 5). A double-sided t-test confirmed that the treatment group wrote texts with a statistically significantly higher quality of formal argumentation: t-value= -3.622 and p<0.001. For the perceived quality of argumentation, we found that on a Likert scale from 1 to 5 points (1: very persuasive, 5: not persuasive) texts from the treatment group achieved an average value of 2.62 (SD= 0.96). Participants using the alternative application wrote texts with a mean value of the perceived quality of argumentation of 3.21 (SD= 1.19). A double-sided t-test showed that the difference was statistically significant: t-value= -2.654 and p-value= 0.0105 (p<0.05). This clearly proves our hypothesis that individual feedback on students' argumentation helps them to write more convincing texts. The results show that students using AL wrote texts with a better formal quality of argumentation as well as a better perceived quality of argumentation compared to the ones using the traditional approach.

Qualitative User Feedback

As described above, we also included open questions in our survey to receive the participants' opinions about the tool they used. The general attitude for AL was very positive. Especially the fast and direct feedback, the graph-like visualization of the argumentation structure and the summarizing scores were mentioned quite often. However, sometimes AL was not correctly classifying claims and premises, which users suggest to improve. We translated the responses from German and categorized the most representative responses in table 6.

DISCUSSION

Our evaluation demonstrated that adaptive and individual feedback on students' argumentation skills helps them to write more persuasive texts. Not only the perceived argumentation quality but also the formal quality of the argumentation was significantly higher for students using AL compared to the ones using the alternative tool. We believe that cognitive dissonance theory explains this effect. The right level of feedback

Group	Feature
On user interaction	*"Easy handling. Especially marked texts with colour and percentage values. Fast reaction of the tool. Motivated to write."*
On writing support	*"I was convinced by the mind map argumentation graphics. If you write longer texts, you can get lost quickly. With this tool you can see how the sentences and argumentation stand together."*
On visualization	*"It was positive that the tool presented which premises support which of my claims and which arguments hung, so to speak, freely in the air."*
On graphics and colour	*"I liked that the tool used colors to highlight the various elements and graphics to represent my text. Furthermore, the percentages of how my text was written were helpful and it is exciting to see how the tool judges my text."*
On speed of the tool	*"Very fast and instant feedback."*
Improvements on feedback accuracy	*"I'm not sure how well this algorithm really understands what I'm writing."*
Improvements on user on-boarding	*"Better introduction would be good, you had to try something before you knew how to do it. Maybe a short YouTube tutorial at the beginning, then you know more about what is important."*
Improvements on the summarizing scores	*"There are only three evaluation points (readability, coherence and persuasiveness) and it is unclear how relevant the individual factors are."*

Table 6. Representative examples of qualitative user responses

on a student's skills, such as argumentation skills, leads to cognitive dissonance and thus to motivation to change the behavior, belief or knowledge to learn how to argue. In order to successfully use a learning tool in a real-world scenario, positive technology acceptance is very important to ensure students perceive the usage of the tool as helpful, useful and easy to interact with. This will foster motivation and engagement to use the learning application. The positive perceived usefulness and intention to use of AL provides very promising results to use it as a feedback application in different learning settings. We believe that the proven short-term improvement on argumentative texts in a possibly continuous use case could lead to cognitive dissonance for the user and motivate him to learn and thus improve his skills based on cognitive dissonance theory [18]. This theory supports our underlying hypothesis that individual and personal feedback on a student's argumentation motivates the student to improve their skill level. Therefore, our work makes several contributions to current research. To the best of our knowledge, this study is one of the first to present evaluated design knowledge on how to build a learning tool to train argumentation skills based on adaptive and intelligent feedback. It provides a basis for researchers who also aim to develop learning tools to train meta cognition skills to compare their solution with ours. Lecturers and educational institutions can now use our design principles and findings to create their own learning tools for providing adaptive and intelligent support of argumentation skills in large-scale scenarios. The main improvement suggestions from users in the qualitative feedback was that the feedback of AL must be as accurate as possible in order for users to be motivated to use the tool. In most of the cases, this seemed to be the case in our experiment, since only about twenty percent of users mentioned this issue. However, the accuracy of our feedback algorithm probably leaves the largest improvement space. We see two main options to tune the performance of our algorithm: a) enrich the corpus with more annotated texts and b) improve the performance of our models to enhance the performance of both the argument component identification classifier and the argument relation identification classifier. We will approach the first point by annotating the written peer reviews which we collected in our experiment, and add them to the corpus. For the second issue, we will experiment with further classification approaches, such as deep learning algorithms or transfer learning models like BERT [12] or ELMO [43].

Moreover, we want to ensure that the three overall scores are more transparent and understandable for the users. Therefore, we will design new calculation models for the readability, coherence and persuasiveness of the text and provide more accurate and transparent action steps on how to achieve a higher rating. In our experiment we prove the short-term influence of AL on a student's argumentation skills. For future work we suggest to measure the long-term learning effects on students' skills. This can be achieved with a longitudinal study in a real-world learning setting, e.g., in supporting the writing of peer reviews in business innovation lectures. Therefore, our next step will be to conduct a field experiment with three groups to evaluate the long-term impact of adaptive and intelligent feedback (provided by our feedback algorithm) on the development of students' argumentation quality. We will rely on one control group (participants will not receive any feedback) and two treatment groups. Participants in treatment group 1 will receive information on how their argumentation quality was scored and general feedback on how to improve it, whereas participants in treatment group 2 will receive information on how their argumentation quality was scored as well as individualized feedback based on their own performance on how they could improve their argumentation quality. The functionalities necessary for the treatments will be implemented into our existing learning system ([47]). At the end of the study, we want to contribute with an evaluated learning tool that can be used in a learning-teaching scenario where students fulfill a certain exercise in a lecture (e.g., writing convincing statements for a business model) and additionally receive feedback on their argumentation on the given text.

CONCLUSION

In this research, we designed, built and evaluated AL, an adaptive IT tool that provides students with feedback on the argumentation structure of a text by leveraging the recent advances of AM algorithms. We compared AL to a proven argumentation writing support approach in a rigorous user study with 54 participants. We found that students using AL wrote more convincing texts with a better formal quality of argumentation compared to the traditional approach. The perceived learning and intention to use provided promising results to use this tool as a feedback application in different learning settings. Our results also offer design suggestions to further improve educational feedback applications based on intelligent algorithms. With NLP and ML becoming more powerful, we hope our work will attract other researchers to design and build more intelligent tutoring systems for other learning scenarios or meta cognition skills.

REFERENCES

[1] Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet Argument Corpus 2.0: An SQL schema for Dialogic Social Media and the Corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Paris, France.

[2] Ritu Agarwal and Elena Karahanna. 2000. Time Flies When You're Having Fun: Cognitive Absorption and Beliefs about Information Technology Usage. *MIS Quarterly* 24, 4 (12 2000), 665. DOI: http://dx.doi.org/10.2307/3250951

[3] Ahmet Aker, Alfred Sliwa, Yuan Ma, Ruishen Lui, Niravkumar Borad, Seyedeh Ziyaei, and Mina Ghobadi. 2017. What works and what does not: Classifier and feature analysis for argument mining. In *Proceedings of the 4th Workshop on Argument Mining*. Association for Computational Linguistics, Copenhagen, Denmark, 91–96. DOI:http://dx.doi.org/10.18653/v1/W17-5112

[4] S. J. Ashford. 1986. Feedback-Seeking in Individual Adaptation : A Resource Perspective. *Academy of Management Journal* 29, 3 (9 1986), 465–487. DOI: http://dx.doi.org/10.2307/256219

[5] Paul Black and Dylan Wiliam. 2009. Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability* 21, 1 (2009), 5–31. DOI: http://dx.doi.org/10.1007/s11092-008-9068-5

[6] Elena Cabrio and Serena Villata. 2014. Towards a Benchmark of Natural Language Arguments. *CoRR* abs/1405.0941 (2014).

[7] Glenn Rowe Chris Reed, Raquel Mochales Palau and Marie-Francine Moens. 2008. Language Resources for Studying Argument. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri (Ed.). European Language Resources Association (ELRA), Marrakech, Morocco.

[8] Mike Cohn. 2004. *User Stories Applied For Agile Software Development*. Technical Report.

[9] Harris M. Cooper. 1988. Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society* 1, 1 (1988), 104–126. DOI: http://dx.doi.org/10.1007/BF03177550

[10] R De Groot, R Drachman, R Hever, B Schwartz, U Hoppe, A Harrer, M De Laat, R Wegerif, B M Mclaren, and B Baurens. 2007. *Computer Supported Moderation of E-Discussions: the ARGUNAUT Approach*. Technical Report. http://www.argunaut.org

[11] Lingjia Deng and Janyce Wiebe. 2015. MPQA 3.0: An Entity/Event-Level Sentiment Corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, 1323–1328.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (10 2018). http://arxiv.org/abs/1810.04805

[13] Rosalind Driver, Paul Newton, and Jonathan Osborne. 2000. Establishing the norms of scientific argumentation in classrooms. *Science Education* 84, 3 (5 2000), 287–312. DOI: http://dx.doi.org/10.1002/(SICI)1098-237X(200005)84:3<287::AID-SCE1>3.0.CO;2-A

[14] Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. 2015. On the Role of Discourse Markers for Discriminating Claims and Premises in Argumentative Discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 2236–2242.

[15] Frans H. van Eemeren, Rob Grootendorst, Ralph H. Johnson, Christian Plantin, Charles A. Willard, Rob Grootendorst, Ralph H. Johnson, Christian Plantin, and Charles A. Willard. 1996. *Fundamentals of Argumentation Theory*. Routledge. DOI: http://dx.doi.org/10.4324/9780203811306

[16] Andrew J. Elliot and Patricia G. Devine. 1994. On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology* 67, 3 (1994), 382–394. DOI: http://dx.doi.org/10.1037/0022-3514.67.3.382

[17] Charles Fadel, Maya Bialik, and Bernie Trilling. 2015. *Four-dimensional education : the competencies learners need to succeed*. 177 pages.

[18] Leon Festinger. 1962. Cognitive Dissonance. *Scientific American* 207, 4 (10 1962), 93–106. DOI: http://dx.doi.org/10.1038/scientificamerican1062-93

[19] Frank Fischer, Ingo Kollar, Karsten Stegmann, and Christof Wecker. 2013. Toward a Script Theory of Guidance in Computer-Supported Collaborative Learning. *Educational psychologist* 48, 1 (1 2013), 56–66. DOI: http://dx.doi.org/10.1080/00461520.2012.748005

[20] J.L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76, 5 (1971), 378–382.

[21] Jürgen Flender, Ursula Christmann, and Norbert Groeben. 1999. Entwicklung und erste Validierung einer Skala zur Erfassung der passiven argumentativ-rhetorischen Kompetenz. *Zeitschrift für Differentielle und Diagnostische Psychologie* 20, 4 (9 1999), 309–325. DOI: http://dx.doi.org/10.1024//0170-1789.20.4.309

[22] R Flesch. 1943. Marks of readable style; a study in adult education. *Teachers College Contributions to Education* 897 (1943), ix + 69–ix + 69.

[23] Pauline Carolyne Fortes and Abdellatif Tchantchane. 2010. Dealing with large classes: A real challenge. *Procedia - Social and Behavioral Sciences* 8 (2010), 272–280. DOI: http://dx.doi.org/10.1016/j.sbspro.2010.12.037

[24] Hansjörg Fromm, Thiemo Wambsganss, and Matthias Söllner. 2019. Towards a Taxonomy of Text Mining Features. In *European Conference of Information Systems (ECIS)*. 1–12.

[25] Jochen. Gläser and Grit. Laudel. 2010. *Experteninterviews und qualitative Inhaltsanalyse : als Instrumente rekonstruierender Untersuchungen*. VS Verlag für Sozialwiss. http://www.springer.com/de/book/9783531172385

[26] Ivan Habernal and Iryna Gurevych. 2015. *Exploiting Debate Portals for Semi-Supervised Argumentation Mining in User-Generated Web Discourse*. Technical Report. 17–21 pages. https://github.com/habernal/emnlp2015

[27] John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of Educational Research* 77, 1 (2007), 81–112. DOI: http://dx.doi.org/10.3102/003465430298487

[28] Chenn Jung Huang, Shun Chih Chang, Heng Ming Chen, Jhe Hao Tseng, and Sheng Yuan Chien. 2016. A group intelligence-based asynchronous argumentation learning-assistance platform. *Interactive Learning Environments* 24, 7 (2016), 1408–1427. DOI: http://dx.doi.org/10.1080/10494820.2015.1016533

[29] David H. Jonassen and Bosung Kim. 2010. Arguing to learn and learning to argue: Design justifications and guidelines. *Educational Technology Research and Development* 58, 4 (2010), 439–457. DOI: http://dx.doi.org/10.1007/s11423-009-9143-8

[30] Barbara Konat, John Lawrence, Joonsuk Park, Katarzyna Budzynska, and Chris Reed. 2016. A Corpus of Argument Networks: Using Graph Properties to Analyse Divisive Issues. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Paris, France.

[31] Timothy Koschmann. 1996. *Paradigm Shifts and Instructional Technology*. Technical Report. 1–23 pages. http://opensiuc.lib.siu.edu/meded_books/4

[32] Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Methodology*. Sage Publications, Inc., Beverly Hills, CA. http://www.amazon.ca/gp/product/0803914989/ref=wl_it_dp/702-0885532-1303250?ie=UTF8&coliid=I3UJ8HY4GH9OWF&colid=1DVGN4EKR6AVM

[33] Klaus Krippendorff. 2004. Measuring the Reliability of Qualitative Text Analysis Data. *Quality and Quantity* 38, 6 (01 Dec 2004), 787–800. DOI: http://dx.doi.org/10.1007/s11135-004-8107-7

[34] Deanna Kuhn. 1992. Thinking as Argument. *Harvard Educational Review* 62, 2 (7 1992), 155–179. DOI: http://dx.doi.org/10.17763/haer.62.2.9r424r0113t670l1

[35] Deanna. Kuhn. 2005. *Education for thinking*. Harvard University Press. 209 pages. http://www.hup.harvard.edu/catalog.php?isbn=9780674027459

[36] Katja Lehmann, Matthias Söllner, and Jan Marco Leimeister. 2016. Design and evaluation of an IT-based peer assessment to increase learner performance in large-scale lectures. In *International Conference on Information Systems*. Association for Information Systems. DOI:http://dx.doi.org/10.2139/ssrn.3159160

[37] Marco Lippi and Paolo Torroni. 2015. Argumentation Mining: State of the Art and Emerging Trends. *IJCAI International Joint Conference on Artificial Intelligence* 2015-Janua, 2 (2015), 4207–4211. DOI: http://dx.doi.org/10.1145/2850417

[38] Jack Mezirow. 1991. *Transformative dimensions of adult learning.* Jossey-Bass, San Francisco, CA 94104-1310.

[39] Raquel Mochales Palau and Aagje Ieven. 2009. Creating an argumentation corpus: do theories apply to real arguments? A case study on the legal argumentation of the ECHR. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Law (ICAIL 2009), Twelfth international conference on artificial intelligence and law (ICAIL 2009)., Barcelona, Spain, 8-12 June 2009*. ACM, 21–30.

[40] E. Michael Nussbaum, Denise L. Winsor, Yvette M. Aqui, and Anne M. Poliquin. 2007. Putting the pieces together: Online argumentation vee diagrams enhance thinking during discussions. *International Journal of Computer-Supported Collaborative Learning* 2, 4 (11 2007), 479–500. DOI: http://dx.doi.org/10.1007/s11412-007-9025-1

[41] OECD. 2018. The Future of Education and Skills - Education 2030. (2018). DOI: http://dx.doi.org/2018-06-15

[42] Jonathan F. Osborne, J. Bryan Henderson, Anna MacPherson, Evan Szu, Andrew Wild, and Shi Ying Yao. 2016. The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching* 53, 6 (2016), 821–846. DOI:http://dx.doi.org/10.1002/tea.21316

[43] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. Association for Computational Linguistics (ACL), 2227–2237. DOI: http://dx.doi.org/10.18653/v1/n18-1202

[44] Jean Piaget, Terrance Brown, and Kishore Julian Thampy. 1986. The Equilibration of Cognitive Structures: The Central Problem of Intellectual Development. Jean Piaget , Terrance Brown , Kishore Julian Thampy. *American Journal of Education* 94, 4 (8 1986), 574–577. DOI:http://dx.doi.org/10.1086/443876

[45] Niels Pinkwart, Kevin Ashley, Collin Lynch, and Vincent Aleven. 2009. *Evaluating an Intelligent Tutoring System for Making Legal Arguments with Hypotheticals*. Technical Report. 401–424 pages. http://iaiedsoc.org/pub/1302/file/19_4_05_Pinkwart.pdf

[46] Eric Ries. 2011. The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses. (2011). https://books.google.de/books?hl=de&lr=&id=tvfyz-4JILwC&oi=fnd&pg=PA15&dq=lean+startup+ries&ots=8H8ay99lrV&sig=NTC5ybgihYWRr6m9aT0XH-F6Ixc#v=onepage&q=leanstartupries&f=falsehttp://en.wikipedia.org/wiki/Lean_Startup

[47] Roman Rietsche, Kevin Duss, Jan Martin Persch, and Matthias Söllner. 2018. Design and Evaluation of an IT-based Formative Feedback Tool to Foster Student Performance Understanding and Designing Trust in Information Systems View project Future of Collaboration View project. In *Thirty Ninth International Conference on Information Systems*. https://www.researchgate.net/publication/329450233

[48] D. Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional Science* 18, 2 (6 1989), 119–144. DOI: http://dx.doi.org/10.1007/BF00117714

[49] Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument Extraction from News. In *Proceedings of the 2nd Workshop on Argumentation Mining*. Association for Computational Linguistics, Denver, Colorado, 56–66.

[50] Katharina Scheiter and Peter Gerjets. 2007. Learner control in hypermedia environments. *Educational Psychology Review* 19, 3 (2007), 285–307. DOI: http://dx.doi.org/10.1007/s10648-007-9046-3

[51] Oliver Scheuer. 2015. *Towards adaptive argumentation learning systems*. https://www.researchgate.net/publication/298087259

[52] Oliver Scheuer, Frank Loll, Niels Pinkwart, and Bruce M. McLaren. 2010. Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning* 5, 1 (2010), 43–102. DOI: http://dx.doi.org/10.1007/s11412-009-9080-x

[53] Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From argumentation mining to stance classification. *NAACL HLT 2015* (2015), 67–77.

[54] Elliot Soloway, Mark Guzdial, and Kenneth E Hay. 1994. Learner-Centered Design The Challenge For WC1 In The Xst Century. *Interactions* (1994), 36–48.

[55] Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. *Applying Argumentation Schemes for Essay Scoring*. Technical Report. 69–78 pages. http://acl2014.org/acl2014/W14-21/pdf/W14-2110.pdf

[56] Christian Stab and Iryna Gurevych. 2014a. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers ,*. 1501–1510. http://www.ukp.tu-darmstadt.de

[57] Christian Stab and Iryna Gurevych. 2014b. Identifying Argumentative Discourse Structures in Persuasive Essays. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)(Oct. 2014), Association for Computational Linguistics*. 46–56. www.ukp.tu-darmstadt.de

[58] Christian Stab and Iryna Gurevych. 2017a. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics* 43, 3 (9 2017), 619–659. DOI:http://dx.doi.org/10.1162/COLI{_}a{_}00295

[59] Christian Stab and Iryna Gurevych. 2017b. *Recognizing Insufficiently Supported Arguments in Argumentative Essays*. Technical Report. 980–990 pages. www.ukp.tu-darmstadt.de

[60] Karsten Stegmann, Christof Wecker, Armin Weinberger, and Frank Fischer. 2012. Collaborative argumentation and cognitive elaboration in a computer-supported collaborative learning environment. *Instructional Science* 40, 2 (2012), 297–323. DOI: http://dx.doi.org/10.1007/s11251-011-9174-5

[61] Pontus Stenetorp, Sampo Pyysalo, and Goran Topi. 2012. BRAT : a Web-based Tool for NLP-Assisted Text Annotation. Figure 1 (2012), 102–107.

[62] Daniel D Suthers and Christopher D Hundhausen. 2001. *European Perspectives on Computer-Supported Collaborative Learning*. Technical Report. 577–584 pages. http://lilt.ics.hawaii.edu/papers/2001/Suthers-Hundhausen-Euro-CSCL-2001.pdf

[63] Stephen E. Toulmin. 1984. *Introduction to Reasoning*.

[64] Viswanath Venkatesh and Hillol Bala. 2008. Technology Acceptance Model 3 and a Research Agenda on Interventions. *Decision Sciences* 39, 2 (5 2008), 273–315. DOI: http://dx.doi.org/10.1111/j.1540-5915.2008.00192.x

[65] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly* 27, 3 (2003), 425–478.

[66] Jan vom Brocke, Wolfgang Maaß, Peter Buxmann, Alexander Maedche, Jan Marco Leimeister, and Günter Pecht. 2018. Future Work and Enterprise Systems. *Business and Information Systems Engineering* 60, 4 (2018), 357–366. DOI: http://dx.doi.org/10.1007/s12599-018-0544-2

[67] Jan vom Brocke, Alexander Simons, Kai Riemer, Björn Niehaves, Ralf Plattfaut, and Anne Cleven. 2015. Standing on the Shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research. *Communications of the Association for Information Systems* 37, 1 (8 2015). DOI: http://dx.doi.org/10.17705/1CAIS.03709

[68] Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014. A Review Corpus for Argumentation Analysis. In *Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 8404 (CICLing 2014)*. Springer-Verlag New York, Inc., New York, NY, USA, 115–127.

[69] Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A Corpus for Research on Deliberation and Debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. Istanbul, Turkey, 812–817. ACL Anthology Identifier: L12-1643.

[70] Thiemo Wambsganss and Roman Rietsche. 2019. Towards Designing an Adaptive Argumentation Learning Tool. International Conference on Information Systems (ICIS).

[71] Armin Weinberger and Frank Fischer. 2006. A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers and Education* 46, 1 (2006), 71–95. DOI:http://dx.doi.org/10.1016/j.compedu.2005.04.003