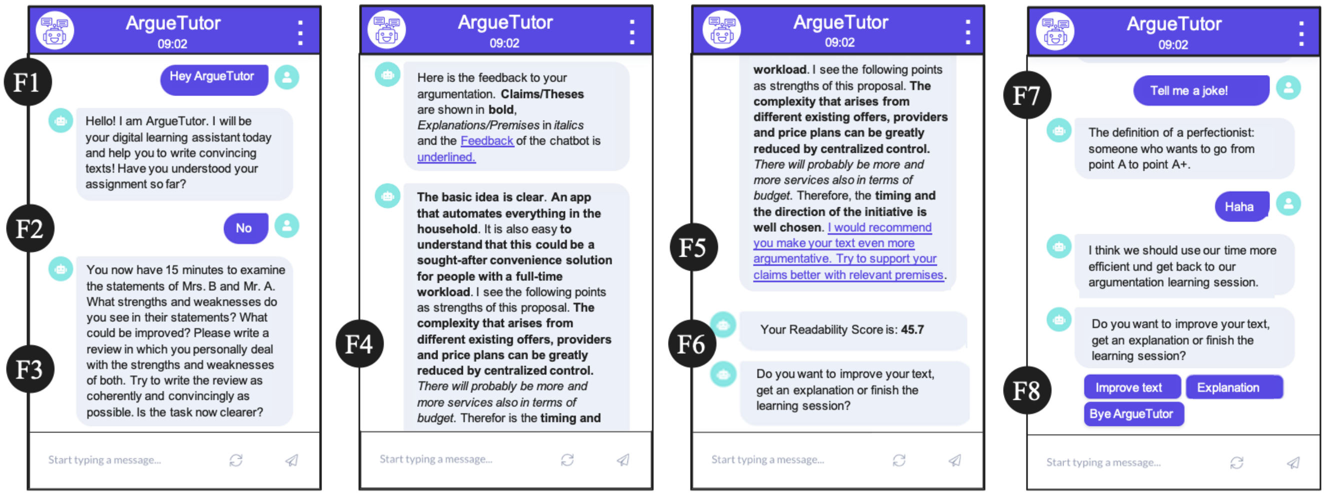
ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills

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

Tobias Küng
tobias.kueng@student.unisg.ch
University of St.Gallen
St.Gallen, Switzerland

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

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



Hey ArgueTutor

Hello! I am ArgueTutor. I will be your digital learning assistant today and help you to write convincing texts! Have you understood your assignment so far?

No

You now have 15 minutes to examine the statements of Mrs. B and Mr. A. What strengths and weaknesses do you see in their statements? What could be improved? Please write a review in which you personally deal with the strengths and weaknesses of both. Try to write the review as coherently and convincingly as possible. Is the task now clearer?

Here is the feedback to your argumentation. **Claims/Theses** are shown in **bold**, *Explanations/Premises* in *italics* and the Feedback of the chatbot is underlined.

The basic idea is clear. **An app that automates everything in the household**. It is also easy **to understand that this could be a sought-after convenience solution for people with a full-time workload**. I see the following points as strengths of this proposal. **The complexity that arises from different existing offers, providers and price plans can be greatly reduced by centralized control.** *There will probably be more and more services also in terms of budget*. Therefor is the **timing and**

workload. I see the following points as strengths of this proposal. **The complexity that arises from different existing offers, providers and price plans can be greatly reduced by centralized control.** *There will probably be more and more services also in terms of budget*. Therefore, the **timing and the direction of the initiative is well chosen**. I would recommend you make your text even more argumentative. Try to support your claims better with relevant premises.

Your Readability Score is: **45.7**

Do you want to improve your text, get an explanation or finish the learning session?

Tell me a joke!

The definition of a perfectionist: someone who wants to go from point A to point A+.

Haha

I think we should use our time more efficient und get back to our argumentation learning session.

Do you want to improve your text, get an explanation or finish the learning session?

Figure 1: Screenshot of our adaptive dialog-based learning system: a user conducts a certain writing exercise and receives adaptive tutoring and feedback on the argumentation quality of her text

ABSTRACT

Techniques from Natural-Language-Processing offer the opportunities to design new dialog-based forms of human-computer interaction as well as to analyze the argumentation quality of texts. This can be leveraged to provide students with adaptive tutoring when doing a persuasive writing exercise. To test if individual tutoring for students' argumentation will help them to write more convincing texts, we developed ArgueTutor, a conversational agent that tutors students with adaptive argumentation feedback in their learning journey. We compared ArgueTutor with 55 students to a traditional writing tool. We found students using ArgueTutor wrote more convincing texts with a better quality of argumentation compared to the ones using the alternative approach. The measured level of enjoyment and ease of use provides promising results to use our tool in traditional learning settings. Our results indicate that dialog-based learning applications combined with NLP text feedback have a beneficial use to foster better writing skills of students.

CCS CONCEPTS

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

KEYWORDS

educational applications, pedagogical conversational agents, argumentation learning, adaptive learning

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05...$15.00
https://doi.org/10.1145/3411764.3445781

ACM Reference Format:
Thiemo Wambsganss, Tobias Küng, Matthias Söllner, and Jan Marco Leimeister. 2021. ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3411764.3445781

1 INTRODUCTION

Today, information can constantly be accessed, so people need to develop skills that go beyond the replication of factual knowledge. Hence, the requirements of job profiles are shifting towards more interdisciplinary, ambiguous and creative tasks [69]. As a result, educational institutions are called to evolve their curricula when it comes to the compositions of skills and knowledge conveyed [65]. Most notably, teaching metacognition skills to students, such as critical thinking, collaboration or problem-solving, have become a central interested of educators [21]. International organizations, such as the Organization for Economic Cooperation and Development (OECD) or the World Economic Forum (WEF), called for a change in the education of students to equip future work forces with the right skill set for digitization [81]. In fact, the OECD included metacognition skills as a major element in their Learning Framework 2030 (OECD, 2018). One subclass of metacognition skills represents the skill of arguing in a structured, reflective and logical form [66]. 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 [36]. Starting with studies from Aristotle, the ability to form convincing arguments is recognized as the foundation for persuading an audience of novel ideas and plays a major role in productive democratic civil discourse, e.g., for citizens to assess if a certain news is fake or not [17]. To develop skills such as argumentation, it is of great importance for the individual student to receive continuous tutoring and feedback throughout their learning journey [5, 28]. Thus, institutions, such as universities, face the challenge of providing individual learning conditions, since every student would need a personal tutor to have an optimal learning environment to learn logical and structured argumentation [70]. However, this is naturally hindered due to traditional large-scale lectures or due to the growing field of distance learning scenarios such as massive open online courses (MOOCs) [56]. The current Covid19-pandemic crisis has strengthened this effect even further, since due to governmental lockdowns distance-learning scenarios have become a reality for many educators.

One possible solution to imitate meaningful, individual instructor–learner interactions are pedagogical conversational agents (PCAs) [84]. PCAs are software programs that communicate with users through natural language interaction interfaces [57, 79]. They have been successfully used to adaptively support learners to conduct a task by mimicking the gold standard of human tutors, e.g., for programming tasks [83], mathematical skills [8] as well as for learning factual knowledge [53]. Due to developments in domains such as Natural Language Processing (NLP) and Machine Learning (ML), PCAs are also becoming increasingly valuable for more creative and harder to grasp skills such as argumentation. Researchers use Argumentation Mining (AM) to develop algorithms that extract and assess the argumentative quality from given texts [40, 72]. This information can be used to score the argumentation level of a text and provide adaptive tutoring and feedback concerning the persuasiveness of a text. Scientists, especially from the field of educational technology, have designed tools to support the active teaching of argumentation to students with input masks, representational guidelines or adaptive writing support systems to enhance students' learning of argumentation (e.g., [14, 45, 49, 73]). By using an adaptive PCA as a tutor for argumentation skill learning, based on recent developments in ML and NLP, students would be able to receive adaptive tutoring during the writing process autonomously and independently of the instructor, time and place [78]. However, current literature sparsely investigates an approach with principles and proof on how to design an adaptive dialog-based learning system that guides and helps students to learn how to argue when writing a persuasive text.

Given this potential for leveraging a conversational learning tool in combination with argumentation mining, we designed and built ArgueTutor (short for *Argumentation Tutor*), an adaptive dialog-based tutoring system that provides students with adaptive and instant feedback, theoretical input and step-by-step guidance during their writing process. We followed two different development approaches: 1) a rigorous theory-motivated approach, where we systematically searched literature in the field of educational technology and HCI following [13, 69] to carefully derive requirements and principles for a design of ArgueTutor, and 2), a user-centered design approach, where we conducted twelve semi-structured interviews with students to derive user stories. Based on these user needs, we built low-fidelity prototypes of ArgueTutor to test different design hypotheses with potential users to learn about the interaction flow of a dialog-based learning tool for argumentation. With these two approaches, we present our final version of ArgueTutor.

To design an adaptive dialog-based learning tool, we a) trained the intents of a conversational tutor based on rigorously scripted teacher-student conversations and b) developed an argumentation mining model to assess the argumentation quality of student texts. To do so, we leveraged the argumentation annotated business-model peer review corpus of [75], since the texts are derived from a pedagogical scenario and built on a rigorous annotation guideline with a moderate agreement. We trained and tuned a transfer learning model based on [16] to classify the argumentation quality of student texts. In combination with the trained chat intents, this model now serves as the underlying adaptive tutoring algorithm of ArgueTutor.

To determine the impact of ArgueTutor on students' argumentation skills and perception during the learning process, we evaluated our learning tool in comparison with a discussion scripting approach that provided general argumentation feedback based on argumentation theory [66], a traditional approach for supporting persuasive writing in large-scale scenarios (e.g., [22, 51, 55]). In a study with 55 students, we observed that participants who used ArgueTutor for a persuasive writing task wrote formally more argumentative texts. Furthermore, the perceived persuasiveness of these texts was significantly higher than of the texts from the traditional tool. We also measured the perceived ease of use and the level of enjoyment of both tools using key constructs [67, 68]. We found that both constructs provide promising results for the usage of ArgueTutor as a standard learning tool in lectures. The results suggest that ArgueTutor 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, with ArgueTutor we introduce the first intelligent dialog-based tutoring system for argumentation skills. Second, we show its effectiveness through rigorously comparing ArgueTutor with a traditional writing support scenario for argumentation skills. The results provide insights into the benefits of leveraging NLP and ML for designing intelligent dialog-based tutoring systems to foster argumentative writing in a student's learning journey. Our results demonstrate an exemplary scenario of supporting metacognition skills in a scalable and individual way in possible large- or distance-scale scenarios. Finally, we provide design knowledge for other researchers and educators to design and compare similar adaptive learning tools to foster the metacognition skill learning of students as a step to contribute to the OECD Learning framework 2030 towards a metacognition-skill-based education.

2 RELATED WORK AND CONCEPTUAL BACKGROUND

Our work was inspired by previous studies on technology-based learning systems for argumentation, by studies about argumentation mining algorithms and by PCAs and the ICAP framework [10], which serves as an underlying conceptual model for our main hypothesis.

2.1 Technology-Based Learning Systems for Argumentation

Argumentation skills build the basis for our daily communication and thinking. In general, argumentation aims at increasing or decreasing the acceptability of a controversial standpoint [20]. Logical, structured arguments are a required precondition for persuasive conversations, general decision-making and drawing acknowledged conclusions. As [35] mentions, the skill to argue is of great significance, not only in professional environments for communication, collaboration and for solving difficult problems but also for most of our daily life. Research calls for logical argumentation support, especially when it comes to democratic civil discourse in which logical argumentation is one of the major elements for efficient and productive civil debates [17]. However, approaches for teaching argumentation are limited. [32] identified three major challenges for teaching it: "*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 [18, 37]. 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. To train argumentation, it is of great importance for the individual student to receive continuous feedback and tutoring throughout her learning journey [28, 70]. Furthermore, even in fields where argumentation is part of the curriculum, such as law or logic, a teacher's ability to teach argumentation is naturally limited by constraints on time and availability [54]. Especially in increasingly common large-scale lectures or distance learning settings such as MOOCs, the ability to support a student's argumentation skills individually is hindered, since for teachers and professors it is becoming increasingly difficult to provide individual tutoring, such as adaptive support for and feedback to a single student [73, 83].

Hence, researchers, especially from the fields of educational technology and HCI, have analyzed how technology-based learning systems can address this gap and enhance students' learning of argumentation. 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 [55]. This has been investigated across a variety of fields, including law [49], science [45, 64], conversational argumentation [14, 78] and business reviews [73]. Different technological approaches have been used in education. Especially intelligent tutoring systems (ITS) and computer-supported collaborative learning (CSCL) [34] are of special relevance for argumentation learning, since argumentative discussions and debates have been identified as a key for collaborative learning settings. Following [55, 73], three different technology-based argumentation learning systems in the field of CSCL and ITS can be distinguished:

- **Discussion scripting approaches:** Students are provided with structured elements for argumentation learning to stimulate interactions based on script theory of guidance [22]. A common approach is to let users choose between predefined argumentation parts (e.g., [31]) or to provide argumentation theory input to support persuasive writing [55, 66].
- **Representational guidance approaches:** Students are supported by providing representations of argumentation structures with the objective to foster individual reasoning, collaboration and learning. A typical example is to help students to represent their argument structure in the form of node-and-link graphs (e.g., [44, 49]).
- **Adaptive support approaches:** Students are provided with pedagogical feedback on their actions with hints and recommendations to encourage and guide future activities in the writing processes. Typical approaches use an automated evaluation to indicate whether an argument is syntactically and semantically correct (e.g., [49, 59, 61, 73]). However, as [73] states that *"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."*.

Our learning tool combines recent advances in NLP, ML and AM to evaluate new forms of human–computer interaction, such as adaptive PCAs, to intelligently tutor students in their individual argumentation learning process, e.g., with adaptive and instant feedback or theoretical input. Therefore, we build on adaptive support approaches to assess the potential of adaptive argumentation skill learning [40, 73]. In fact, the application of AM and adaptive dialog-based tutoring systems has been motivated but rarely been investigated with design knowledge and empirical evaluation [30, 76, 78]. We compare our tool against a nonadaptive static discussion scripting approach, since it is widespread and a common approach to foster students' argumentation skills in collaborative learning scenarios (e.g., [22, 51].

2.2 Argumentation Mining for Adaptive Learning Systems

The foundation of argumentation mining (AM) is argumentation theory. Argumentation theory is about analyzing the structure and the connection between arguments. One of the most prominent argumentation models is the Toulmin model [66]. Accordingly, 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 [66]. Claim and premise represent the argument components. The claim is the central statement of an argument, representing a controversial text unit. The premises are propositions that either support or attack the claim, underpinning its plausibility. AM, a research field in Computational Linguistics, aims at automatically identifying arguments in unstructured texts [40]. It is gaining momentum in a lot of areas, including the legal domain [43], newswire articles [11, 15], user-generated web content [27, 71] and online debates [7, 19]. During the identification of these argumentation structures, three subtasks can be distinguished:

(1) **identification of argumentative text** paragraphs,
(2) **classification of argumentative text** into claims and premise and
(3) **identification of relationships** between pairs of argument components.

Researchers have developed increasing interest in intelligent writing assistance [58, 60], since it enables adaptive argumentative writing support with tailored feedback about arguments in texts [40, 73]. However, the complexity of using this technology for an adaptive dialog-based learning system has been poorly assessed so far [40]. In our approach we focus on the first two subtasks to assess the argumentation level of a student to provide individual tutoring.

2.3 Pedagogical Conversational Agents and the ICAP Framework to Foster Interactive Learning

Conversational Agents (CA) are software programs that are designed to communicate with users through natural language interaction interfaces [57, 85]. A CA in education is a special form of learning application that interacts with learners individually [30]. We refer to a CA embedded in a pedagogical scenario as a Pedagogical Conversational Agent (PCA). The development of PCAs dates back to the 1970s research stream of Intelligent Tutoring Systems (ITS) [3, 63]. Similar to a human tutor, these systems can present instructions, ask questions [80] and provide immediate feedback [38]. ITS evolved from abstract entities with limited technological possibilities to systems that are able to interact with learners using multiple channels of communication, exhibit social skills and perform different roles, such as tutors [46], motivators or learning companions [30] as well as conducting course evaluations to assist teachers [80].

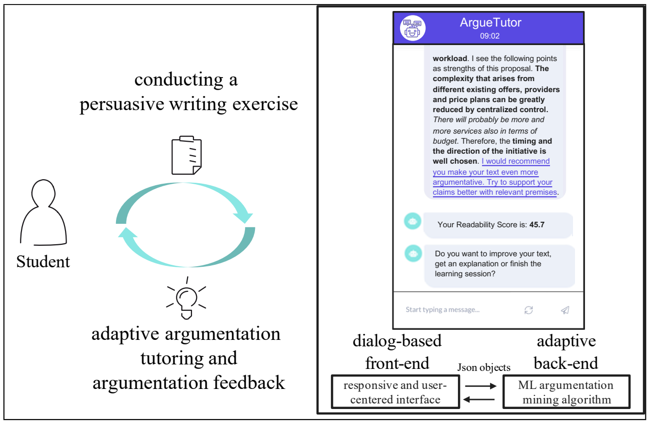
A benefit of using the technology of PCAs compared to traditional technology-enhanced argumentation learning systems is the increasing engagement of the students due to the dialog-based interaction of the learners with the PCA. According to the ICAP Framework (i.e., Interactive, Constructive, Active and Passive Framework) by [10], the learners' engagement with learning materials can range *"from passive to active to constructive to interactive"* [10] and will result in an improved learning outcome. Whereas in passive engagement students only consume or receive learning materials (e.g., reading a text), in active engagement students actively manage the content presentation (e.g., by highlighting important text paragraphs). In the two most engaging forms of interaction according to [10], students deepen their interaction, e.g., by comparing the learning materials with prior knowledge (constructive engagement), by debating with others or asking and answering questions (interactive engagement). Each mode of the ICAP framework corresponds to different types of behaviors and knowledge change processes predicting different learning outcomes [10]. Following this hypothesis, adaptive and interactive dialog-based learning systems are capable of fostering the students' engagement as they add the new component of dialoging to technology-based argumentation learning systems. Compared to common argumentation learning systems, PCAs are able to discuss and tutor students through the learning content – just like human instructors would do. In fact, the successful application of PCAs to adaptively meet individual needs of learners and to increase their learning outcomes has been demonstrated for learning various skills, such as for problem-solving skills [83], programming skills [29], mathematical skills [8] as well as for learning factual knowledge [53], but it has not been investigated for argumentation skills. Therefore, we believe that a theory-motivated and user-centered design of an adaptive PCA combined with intelligent algorithms to provide learning tutoring for argumentation skills by individually assisting students in writing persuasive texts would interactively foster learning according to the ICAP Framework.

3 DESIGN OF A DIALOG-BASED LEARNING SYSTEM

In this section, we will explain how we designed and built the two main components of ArgueTutor: the dialog-based user interaction and the adaptive feedback algorithm in the back end. The basic user interaction concept of ArgueTutor is illustrated in Figure 2. T user does a persuasive writing task and receives adaptive tutoring and feedback on the argumentation.

3.1 Dialog-Based User Interface of ArgueTutor

Deriving theory requirements. To build a theory-motivated and user-centered learning tool, we followed two different approaches: a rigorous theory-driven approach and an agile user-centered approach following the build-measure-learn paradigm [50]. For the rigorous theory-driven approach, we followed the approaches of [13] and [69] to conduct a systematic literature review with the aim of deriving a set of theory requirements for the design of a dialog-based argumentation learning system. We initially focused our research on studies that demonstrate the successful implementation of learning tools for argumentation skills and PCAs. The design of a conversational learning tool for argumentation skills is a complex project that is studied by psychologists, pedagogues and computer scientists with different methods. Therefore, we firstly concentrated on two main literature streams for deriving requirements: educational technology and HCI. We only included studies that dealt with or



workload. I see the following points as strengths of this proposal. **The complexity that arises from different existing offers, providers and price plans can be greatly reduced by centralized control.** *There will probably be more and more services also in terms of budget*. Therefore, the **timing and the direction of the initiative is well chosen**. I would recommend you make your text even more argumentative. Try to support your claims better with relevant premises.

Your Readability Score is: **45.7**

Do you want to improve your text, get an explanation or finish the learning session?

Figure 2: Basic user interaction concept of ArgueTutor: a student is adaptively tutored through a writing exercise with individual argumentation feedback

contribute to a kind of learning tool in the field of argumentation learning or dialog-based learning systems, such as an established pedagogical theory. On this basis, we selected 85 papers for more intensive analysis. We have summarized similar topics of these contributions as literature issues and formed five clusters from them, which served as theory requirements for dialog-based learning tools for metacognition skills.

Deriving user requirements. Besides the rigorous theory-driven approach, we followed a continued user-centered design approach at the same time. As a start, we conducted twelve user interviews with students to receive an initial understanding of the needs and requirements of learners for a dialog-based learning tool for argumentation. Therefore, we followed the expert interview method of [26]. The interview guideline consists of 44 questions and each interview lasted around 30 to 45 minutes. The interviewees were students at our university who are all potential users of an adaptive dialog-based learning tool for metacognition skills. The interviewed students were between 21 and 27 years old; nine were enrolled in a master’s program and three in a bachelor program at our university. All participants were business students; six were male, six were female. We recorded all interviews on audio and transcribed them. Based on the transcription we formatted abstract categories and identified 45 unique user-stories in the twelve interviews. The coding was performed using open coding to form a uniform coding system during evaluation [26]. Based on these results, we gathered user stories and aggregated the most common ones following [12]. Building on the user stories, we designed low-fidelity prototypes of ArgueTutor to test different design hypotheses with end users to learn more about the conversational flows and the human-computer interaction of an adaptive tutoring system for argumentation skills. We started the testing with seventeen low-fidelity paper prototypes and later two digital mock-ups of ArgueTutor. For example, we hypothesized that students aim to learn with a humanized conversational learning tool. We tested this hypothesis with three different paper prototypes. Prototype one was embedded in a rather functional design without incorporating humanized design elements such as a profile picture. Prototypes two and three were designed with a higher level of anthropomorphic elements (e.g., a profile picture and emoticons in a more colloquial conversation). The tutoring and feedback algorithm was simulated by a human. The hypothesis was validated with 16 users. However, we learned that the majority of students rather like a functional design with a low level of anthropomorphic elements. Therefore, the final version of ArgueTutor contributes to that with a rather functional design.

Deriving design principles. In total, we conducted three cycles with a total of 77 different users (16 in cycle one, 16 in cycle two and 45 in cycle three). 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 five design principles on how to build an adaptive dialog-based learning system for argumentation skills illustrated in Table 1. The design principles were followed in the instantiation of our current version of ArgueTutor.

	Design Principle
1)	To design a dialog-based learning system for argumentation skills, provide guidance and support for argumentative writing embedded within conversational elements and an adaptive argumentative text feedback function to allow students to learn interactively.
2)	To design a dialog-based learning system for argumentation skills, employ a web-based conversational agent with a responsive, simple and functional UX to allow students to intuitively use the tool for learning tasks without any distraction.
3)	To design a dialog-based learning system for argumentation skills, provide proactive, individual argumentative tutoring and guidance with explanations based on an argumentation theory to allow students to receive theory-based support whenever they need.
4)	To design a dialog-based learning system for argumentation skills, provide an adaptive feedback function for argumentative texts with an analysis of individual argumentative components and an individual feedback message to allow students to assess their individual argumentation level anytime.
5)	To design a dialog-based learning system for argumentation skills, employ a casual chat function with jokes or fun facts to allow students to take a break from the primary learning activity, but motivate them to continue their learning journey.

Table 1: Derived design principles on how to build an adaptive dialog-based argumentation learning system

User Interaction of ArgueTutor. Following the design principles, ArgueTutor is built as a responsive web-based application that can be used on all kinds of devices. A screenshot of ArgueTutor and its core functionalities (e.g., *F1 - F7*) can be seen in Figure 1 [1]. ArgueTutor consists of two modes: an adaptive learning mode and a casual chat mode. The dialog flow of ArgueTutor with the two interaction modes can be seen in Figure 3. The basis dialog flow of the adaptive learning mode of ArgueTutor was designed according to the didactical learning phases of *"Motivation, Difficulty, Solution, and Practice"* according to [52].

[1] ArgueTutor was designed in German to provide German students with feedback on German texts. However, for ease of understanding our study, we translated parts of the user interface into English (e.g., see Figure 1).

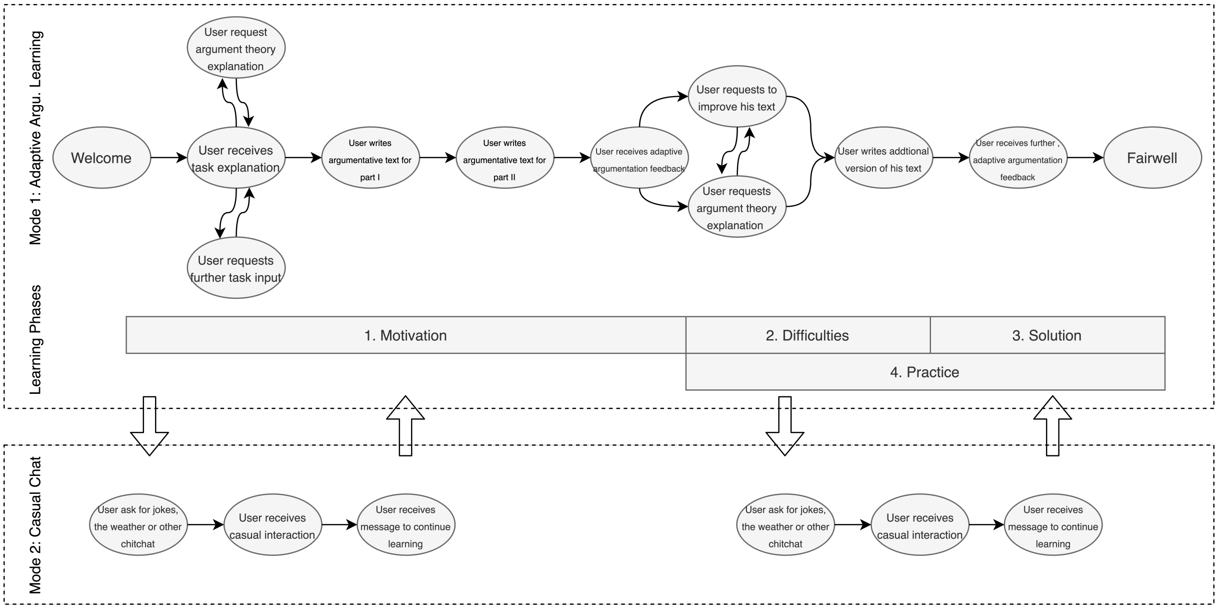

Mode 1: Adaptive Argu. Learning
User request argument theory explanation
Welcome
User receives task explanation
User requests further task input
User writes argumentative text for part I
User writes argumentative text for part II
User receives adaptive argumentation feedback
User requests to improve his text
User requests argument theory explanation
User writes addtional version of his text
User receives further , adaptive argumentation feedback
Fairwell
Learning Phases
1. Motivation
2. Difficulties
3. Solution
4. Practice
Mode 2: Casual Chat
User ask for jokes, the weather or other chitchat
User receives casual interaction
User receives message to continue learning
User ask for jokes, the weather or other chitchat
User receives casual interaction
User receives message to continue learning

Figure 3: The dialog flow of ArgueTutor consists of an adaptive argumentation learning mode (mode 1) and a casual chat function (mode 2)

ArgueTutor guides students through a writing exercise with the aim to imitate a human educator (F1, F2 and F3). The PCA proactively explains a writing task (e.g., students have to write persuasive peer feedback to a fellow student) (F3) and provides hints and explanations when the user asks for help such as argumentation theory input [66]. Moreover, ArgueTutor is always able to provide individual feedback on the argumentation skill level of written student texts by highlighting argumentative components such as claims and premises (F4) and by tutoring students with an individual feedback message on what to improve depending on the student's skill level (F5). Inspired by [42], claims are displayed in bold font, whereas premises are displayed in italic style (F4). Non-argumentative text paragraphs are not highlighted. Additionally, ArgueTutor provides students with an individual summarizing feedback based on the number of premises and claims in the message (F5). For example, if the message contains less than two premises or contains more claims than premises, the user receives a corresponding feedback indicating that the argumentation could be improved with certain improvement suggestions. Besides, based on user feedback, we embedded a readability score in the feedback to provide students with an overview of the quality of their general writing (F6). Therefore, we calculate the readability of the student's text based on the Flesch Reading Ease score [24] and provide a small explanation for the individual score (e.g., if less than 40, recommendations to improve the readability). The interactions between the user and ArgueTutor are mixed between both typing and button selections. While writing a persuasive text for an exercise is typing based, selecting from multiple choices, asking for a task explanation or theory input are button based. Predefined answer buttons help the user to receive an overview of the learning process and ensure both flexibility and efficiency regarding user interactions with ArgueTutor (F8). Moreover, ArgueTutor incorporates a user-initiated and rule-based casual chat mode. Students can ask ArgueTutor to tell jokes, fun facts or talk about the weather to take a break from the primary learning activity (see Figure 1, F7) and thus change from the argumentation learning mode to the casual chat mode at any time they want. To imitate the students having a personal learning session with a human educator, we incorporated several more functions to incorporate real-world conversational elements into ArgueTutor's design. For example, we provided a wide variety of different responses to common conversation states such as "how are you?" as well as positive reinforcement feedback that is typical of a study partner. The result of the dialog path of ArgueTutor from a user perspective is depicted in Figure 3.

The learning mode is based on a rule-based chat system combined with a supervised argumentation mining model to assess the argumentation level of students. The adaptivity of the argumentation feedback is implemented by training a model based on a corpus of persuasive student-written texts. The argumentation theory guidance, the task explanation and other tutoring functions are implemented through rule-based trained chat intents based on a word-to-vec model following the architecture of rasa nlu and rasa core [6].

All in all, we only included simple design elements in the design of ArgueTutor, since we received the user feedback to offer a more formal and functional tutoring experiences that does not distract students. However, we believe these design principles might change depending on culture and context and can thus be easily adapted, e.g., by embedding ArgueTutor in a human persona with a name, face, picture and the use of emojis.

3.2 Argumentation Mining Algorithm of ArgueTutor

To design an adaptive dialog-based learning tool with individual and adaptive guidance, we trained and tuned a transfer learning model that fulfils the users' requirements to receive adaptive and instant tutoring and feedback on their texts.

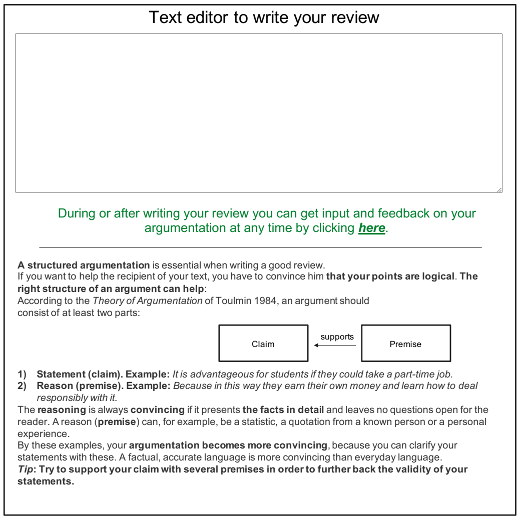
A major prerequisite for developing supervised ML models based on NLP that are able to identify argumentative texts and argument components in written texts is the availability of annotated corpora [25, 39, 74]. We searched the literature for a corpus that fulfilled the following criteria: 1) the corpus contains annotated persuasive student essays, 2) it has a sufficient corpus size to be able to use the trained model in a real-world scenario that fulfills the user requirements and 3) the annotations are based on a rigorous annotation guideline for guiding the annotators towards a moderate agreement. The German business model peer review corpus published in [75] fulfilled all these requirements. The corpus consists of 1000 business model peer feedback essays written by students extracted from a large-scale lecture scenario. The texts are annotated for their argumentative components (claim, premise) as well as for the relations of the components.

Guided by recent literature about AM [40, 72] and Deep Learning [39], we trained a predictive model following the architecture of Bidirectional Encoder Representations from Transformers (BERT) proposed by [16]. Hence, we classified text tokens as claim, premise or as non-argumentative following the argumentation theory of [66]. We used the BERT model from *deepset*[2], since it is available for German and provides a deep pretrained model that was unsupervised while training on domain-agnostic German corpora (e.g., the German Wikipedia). The novelty of this architecture is the ability to capture semantic information from pretrained texts, which can then be used for other downstream tasks without the need for retraining, e.g., for identifying argumentative components. For applying the model, the corpus texts were split into word tokens to fulfill the preparation requirements for BERT. The special preprocessing for BERT was conducted by utilizing the tokenizer and processor provided by deepset and by utilizing spacy[3]. The goal of our model is to provide accurate predictions to identify and classify argument components that can be used for accessing the skill level of students and thus provide adaptive guidance and feedback on how to improve their argumentation.

We split the data into 70 % training, 20 % validation and 10 % test data [4]. For the proposed architecture, the input and output are adapted to the sequence classification task of argument component identification. The last hidden layer is a Recurrent Neural Network with 512 nodes that takes the BERT output and learns to feed into a sigmoid layer that classifies each token according to the predicted label. The proposed model was fine-tuned in several iterations and the best-performing set of hyperparameters included a learning rate of 5e^{-5}, a warm-up and embedding dropout probability of 0.1 and 0.15 respectively. After several iterations, our final BERT model reached a macro f1 score of 73 % for classifying text tokens into claim, premise or non-argumentative tokens.An f1-score of 73 % is a satisfying result when comparing to other studies on student-written argumentation identification. For example, [60] reached an f1-score of 73 % for argumentation stance classification of student-written text in English, and [73] reached an f1-score of 65.4 % for argumentation component classification of German student-written texts. Moreover, we benchmarked our BERT model against bidirectional Long-Short-Term-Memory-Conditional-Random-Fields classifiers (BiLSTM-CRF), since several other authors reached satisfying results with a BI-LSTM-CRF for argumentation compound classification (e.g., [9]). In combination with the corresponding embeddings vocabulary (GloVe) [48] our LSTM only reached an unsatisfying f1 score of 57 % (16 % lower then the 73 % of our BERT model).

3.3 Alternative Argumentation Learning Tool

To evaluate ArgueTutor, we compared it with a nonadaptive discussion scripting application, since this approach was empirically proven to foster the formal quality of the argumentation of students [55, 62]. To control the differences and similarities in the design between the alternative tool and ArgueTutor, we also built the discussion scripting approach ourselves. The learning tool supports the writing process of users with theory-based argumentation input and general nonadaptive theory-based recommendations on how to improve the argumentation level. Users can access the theory input and the general writing recommendations by clicking the "here" button. The theory input and argumentation writing recommendations are then displayed next to the writing editor. To ensure that ArgueTutor and the argumentation writing system are consistent with each other, there are many functionalities that are shared between them. First, the introduction text is the same across both apps. Moreover, the theory input and the general argumentation recommendations are the same for both learning tools following the Toulmin model [66].



Text editor to write your review

During or after writing your review you can get input and feedback on your argumentation at any time by clicking ***here***.

A structured argumentation is essential when writing a good review.
If you want to help the recipient of your text, you have to convince him **that your points are logical**. **The right structure of an argument can help**:
According to the *Theory of Argumentation* of Toulmin 1984, an argument should consist of at least two parts:

1) **Statement (claim). Example:** *It is advantageous for students if they could take a part-time job.*
2) **Reason (premise). Example:** *Because in this way they earn their own money and learn how to deal responsibly with it.*

The **reasoning** is always **convincing** if it presents **the facts in detail** and leaves no questions open for the reader. A reason (**premise**) can, for example, be a statistic, a quotation from a known person or a personal experience.
By these examples, your **argumentation becomes more convincing**, because you can clarify your statements with these. A factual, accurate language is more convincing than everyday language.
***Tip*: Try to support your claim with several premises in order to further back the validity of your statements.**

Figure 4: Basic user interaction concept of alternative argumentation learning tool: students receive theory input and general recommendations on the argumentation of a given text

[2] https://github.com/deepset-ai/FARM
[3] https://spacy.io/

4 EXPERIMENTAL SETUP

In this section, we describe the experimental setup for our study. Our goal was to evaluate our hypothesis that adaptive tutoring on student's argumentation will help them to write more convincing texts. To evaluate our hypothesis, we designed an experiment in which participants were asked to write a peer review based on a given essay. Participants were randomly assigned to a treatment and a control group. The treatment group used ArgueTutor to do the writing exercise, while participants in the control group used the alternative learning tool. We recruited 55 students from our university to take part in our experiment. The experiment was conducted as a web experiment facilitated by the behavioral lab of our university. After randomization, we counted 31 valid results in the treatment and 24 in the control group. Participants of the treatment group had an average age of 22.75 (SD= 1.96), 17 were male, 14 female. In the control group, participants' average age was 23.52 (SD= 2.81), 11 were male, 13 female. All participants were compensated with an equivalent of about 12 USD for a 25- to 30-minute experiment.

4.1 Design and Procedure

The experiment consisted of three main parts: 1) pretest, 2) writing exercise and 3) posttest. The pre- and post-phases were consistent for all participants. In the writing phase, the treatment group used ArgueTutor to conduct a persuasive writing exercise, whereas participants of the control group conducted the same exercise using the alternative tool.

1) Pretest: The experiment started with a pre-survey of 14 questions. Here, we tested three different constructs to assess whether the randomization was successful. First, we asked four items to test the personal innovativeness in the domain of information technology of the participants following [1]. Exemplary items were *"I like to experiment with new information technologies"* or *"If I heard about a new information technology, I would look for ways to experiment with it,"*. Second, we tested the construct of feedback-seeking of individuals following [2]. 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 disagree to 5: totally agree, with 3 being a neutral statement). Third, we controlled for the argumentative competencies of the participants, since we later measured the formal and perceived quality of the argumentation of the written texts. Therefore, we captured the construct of passive argumentative competency following [23], as it is a proven construct to measure argumentative competencies in German. Participants were asked to read a discussion between two teachers concerning the topic "*Does TV make students aggressive?*". We retrieved the entire discussion with the topic and the measurements from [23]. 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*?" [23]. 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 assessed by calculating a score from 0 to 27 following the measurements of [23].

2) Writing exercise: In the writing part of the experiments, we asked the participants to conduct a persuasive writing tasks, simulating a typical student essay homework. We asked the students to write a review about the discussion from the pre-survey. Therefore, students were asked to assess 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 ArgueTutor 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 ArgueTutor were adaptively tutored through the writing exercise, e.g., through theory input, individual recommendations and adaptive argumentation feedback based on our feedback algorithm. Participants using the reference tool retrieved help based on argumentation theory input and general argumentation recommendations during the writing process following [66].

3) Posttest: In the post-survey, we measured the perceived level of enjoyment of the students, since enjoyment during a learning process has a major influence on the adoption of IT tools [41] and on the learning success of students [47]. Therefore, we asked the students the following items: *"The interaction with the learning tool was exciting"* and *"It is fun to interact with the learning tool"* [33]. Moreover, we measured the perceived ease of use of the particpants following the technology acceptance model of [67, 68] and captured the demographics. In total, we asked 13 questions. Finally, we asked 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?*"

4.2 Measurement of Argumentation Quality

Besides measuring the ease of use and the level of enjoyment, 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 [82]. This annotation scheme was applied in various studies and has proven high objectivity, reliability and validity (e.g., [62]). 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 [82]. Therefore, we trained three annotators based on the 15-page annotation guideline of [73] to assess the argumentation components of persuasive reviews. 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 [62], 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 the same three annotators. The objective was to subjectively judge how persuasive the given argumentation is on a Likert scale from 1 to 5 points (1: not very persuasive, 5: very persuasive). We took the mean of all three annotators as a final variable for the formal and the perceived quality of argumentation of the texts.

5 EVALUATION AND RESULTS

To evaluate our hypothesis that adaptive tutoring on students' argumentation will help them to write more convincing texts, our objective was to answer two research questions (RQ):

RQ1: *How effective is ArgueTutor at helping users to write more persuasive texts compared to the traditional approach?*

RQ2: *Do students perceive the interaction with ArgueTutor as easy and enjoyable to use during their writing process, and would they continue to use it in the future?*

To evaluate our first research question, we compare the formal quality of argumentation and perceived quality of argumentation between the written text of the treatment and the control group. Therefore, we applied a *Welch Two Sample t test* to evaluate whether the means of the constructs are significantly different between the groups.

The second research question will be answered by comparing the constructs of perceived ease of use and level of enjoyment for participants using ArgueTutor compared to participants using the alternative tool. We performed a *Welch Two Sample t test* to assess whether differences between both groups are statistically significant. Moreover, we compared the results of ArgueTutor to the midpoints scale to validate a general positive technology acceptance as done in [73]. To ensure that the randomization resulted in randomized groups and to control for potential effects of interfering variables with our small sample size, we compared the differences in the means of the three constructs included in the pretest. 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 groups (for personal innovativeness p= 0.1436, for feedback-seeking of individuals p= 0.7537 and for passive argumentative competency p= 0.8495). This demonstrated that no significant difference in the mean values for these three constructs exists between the groups.

5.1 Argumentation Quality of Written Texts

Group	Formal argumentation	Perceived argu.
Mean ArgueTutor	3.56	3.48
Mean reference tool	2.64	2.80
SD ArgueTutor	1.81	0.83
SD reference tool	1.21	1.43
p value	0.03459	0.03961

Table 2: Results of formal and perceived quality of argumentation between both tools

The mean number of arguments in the texts from participants using ArgueTutor for the writing exercise was 3.56 (SD= 1.81). For the texts from participants using the alternative static tool, we counted a mean of 2.64 arguments (SD= 1.21) (see Figure 2). A double-sided t test confirmed that the treatment group wrote texts with a statistically significantly higher quality of formal argumentation: t value= 2.1738 and p value= 0.03459 ($p<0.05$). For the perceived quality of argumentation, we found that on a Likert scale from 1 to 5 points (1: not very persuasive, 5: very persuasive) texts from the treatment group achieved an average value of 3.48 (SD= 0.83). Participants using the alternative application wrote texts with a mean value of the perceived quality of argumentation of 2.80 (SD= 1.43). A double-sided t test showed that the difference was statistically significant: t value= 2.114 and p value= 0.03961 ($p<0.05$). This clearly proves our hypothesis that adaptive dialog-based tutoring during students' argumentative writing process helps them to write more convincing texts. The results show that students using ArgueTutor wrote texts with a better formal quality of argumentation and with a better perceived quality of argumentation compared to the ones using the static traditional approach.

5.2 Students' Perception of Ease of Use and Enjoyment

Group	ease of use	level of enjoyment
Mean ArgueTutor	3.73	3.41
Mean reference tool	3.45	3.00
SD ArgueTutor	0.64	0.89
SD reference tool	0.69	0.88
p value	0.0286	0.02135

Table 3: Results of the perceived ease of use and level of enjoyment ArgueTutor and the reference tool on a 1 - 5 Likert Scale

To evaluate the students' perception, we calculated the means of the perceived ease of use and the level of enjoyment. We compared the results of ArgueTutor with the results of the alternative tool. The perceived ease of use of ArgueTutor was rated with a mean value of 3.73 (SD= 0.64) and the average of perceived level of enjoyment of ArgueTutor was 3.41 (SD= 0.89). These values are significantly better than the results of the alternative approach. For perceived ease of use we observed a mean value of 3.45 (SD= 0.69) and for perceived level of enjoyment the value was 3.00 (SD= 0.88) for participants from the control group. The results show that the participants of our experiment rated the ease of use of ArgueTutor as an adaptive dialog-based tutoring system 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 3). Moreover, the mean values of ArgueTutor are also very promising when comparing the results to the midpoints. All results are better than the neutral value of 3, indicating a very positive value for level of enjoyment and perceived ease of use. A high level of enjoyment and ease of use is especially important for learning tools to ensure students are experiencing joy in the usage of the tool and they find it easy to interact with. This will foster motivation, engagement and adoption to use the learning application.

5.3 Qualitative User Feedback

We also asked open questions in our survey to receive the participants' opinions about the tool they used. The general attitude for ArgueTutor was quite positive. Participants positively mentioned the fast and adaptive feedback (F4 and F5), the simple conversational interaction flow and the adaptive feedback message with the readability score (F5 and F6) several times. However, participants also asked if ArgueTutor could provide even more detailed feedback about the argumentation (e.g., through displaying argumentative relations between the components) and provide transparent explanations on how the feedback mechanism works. We translated the responses from German and clustered the most representative responses in Table 4.

Cluster	Feature
On the user interaction	*"I liked the interactive dialog-based interaction. It made the task more interesting compared to traditional writing tools with feedback. In a way, you don't feel alone when writing your essay and you get a second try for improvements."*
On feedback accuracy	*"The feedback was pretty accurate and criticized the points I would have criticized myself."*
On readability score and feedback message	*"The readability score and the recommendation seems to be correct and helped me to reflect on my argumentation."*
On the feedback visualization	*"I found it very revealing that different fonts were used to show how my argumentation is structured. This function creates added value."*
On speed of the tool	"*Very fast reaction time of ArgueTutor."*
Improvements on transparency	"*It was not quite clear to me how the feedback algorithm worked and what the score exactly measures."*
Improvements on feedback granularity	*"The feedback regarding my text could be more precise and detailed."*

Table 4: Representative examples of qualitative user responses

6 DISCUSSION

Our research study illustrated that adaptive dialog-based tutoring on students' argumentation skills during a writing exercise helps them to write more persuasive texts. The perceived and the formal argumentation quality was significantly higher for students using ArgueTutor compared to the ones using an alternative discussion scripting apprach for the exact same writing exercise. We believe that the ICAP framework could explain our results. Accordingly, a dialog-based interactive learning journey increases the engagement of the students compared to an active learning scenario. Therefore, the user-centered and theory-based design of an adaptive PCA combined with intelligent algorithms interactively fosters learning according to the conceptual model of the ICAP framework [10]. Also, the perceived ease of use and the level of enjoyment showed positive results for the usage of a learning tool in a real-world scenario. Especially a high level of enjoyment during the learning process is important for the long-term adaption of such learning tools, since this is proven to foster motivation and engagement in the learning process.

Hence, our work makes several contributions to current research. To the best of our knowledge, this study is one of the first to present empirical insights into how to design a dialog-based learning tool to foster argumentation skills of students based on adaptive and intelligent tutoring. It provides a foundation for researchers who also aim to develop learning tools to train metacognition skills to compare their solution with ours (e.g., for empathy skills [77]). Educators can now use our design findings and principles to build their own adaptive PCAs to support argumentation learning in their large-scale or distance-learning scenarios.

Based on the qualitative user feedback, we see three main improvements to enhance ArgueTutor. First, we aim to improve the details of our argumentation text feedback by displaying the argumentative relations between the components through a graph engine. Therefore, students receive an additional overview of the discourse of their argumentation. Second, we aim to improve the transparency of our recommendations and explanations by providing a question mark button in the upper right corner of our PCA. Students can see transparent explanations for the adaptive tutoring at any time they want. Third, we want to train our PCA on different argumentation mining corpora to extend the contribution of our learning tool to other domains and languages (e.g., English student essays or English law cases).

Our study also faces limitations. For the aim of this study, we focused our research on students from our university. Even though it is reasonable to assume that the transferability to other cases is possible without major changes, we cannot prove it with our research design. Moreover, even if 85% of the participants stated they have used a conversational agent before, novelty effects of students using our PCA for the first time cannot be expelled in our empirical results. In our experiment we prove the short-term influence of ArgueTutor 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 tutoring the writing process of peer reviews in business model lectures. Therefore, particularly for analyzing the long-term effect of using ArgueTutor, we aim to implement the artifact into our existing learning management system (blinded for review) and measure long-term effects on usability and the acceptance of skill learning with a PCA during the complete three month life cycle of a lecture. We want to investigate the hypothesis that adaptive dialog-based tutoring influences the long-term argumentation skills of students. 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 do a certain writing exercise and receive adaptive tutoring during the writing process by an intelligent PCA. Regarding the implementation of our PCA, we clearly do not want to replace human tutors, since we believe that skilled teachers will always be able to provide better adaptive skill tutoring than a PCA. However, we hope through our system human tutors can focus more on detailed questions and can devote more time to difficult cases.

7 CONCLUSION

In our research project, we designed, built and evaluated ArgueTutor, an adaptive dialog-based learning system that individually tutors students with task explanations, theory input, guidance and adaptive feedback on the argumentation structure of a text by leveraging the recent advances of AM algorithms. We compared ArgueTutor to a discussion scripting approach in an experiment with 55 participants. We found that students using ArgueTutor to conduct a writing exercise wrote more convincing texts with a better formal quality of argumentation compared to the traditional approach. The perceived ease of use and enjoyment offers promising results to use ArgueTutor as a learning tool in different learning scenarios. All in all, our research offers design knowledge to further improve dialog-based tutoring systems based on techniques from NLP and ML. With further advances of these technologies, we hope our work will attract researchers to design more intelligent tutoring systems for other learning scenarios or metacognition skills and thus contribute to the OECD Learning framework 2030 towards a metacognition-skill-based education.

REFERENCES

[1] 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. https://doi.org/10.2307/3250951

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

[3] R. C. Atkinson and R. M. Shiffrin. 1968. Human Memory: A Proposed System and its Control Processes. *Psychology of Learning and Motivation - Advances in Research and Theory* 2, C (1968), 89–195. https://doi.org/10.1016/S0079-7421(08)60422-3

[4] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. Vol. 43. 479 pages. https://doi.org/10.1097/00004770-200204000-00018

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

[6] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open Source Language Understanding and Dialogue Management. (12 2017). http://arxiv.org/abs/1712.05181

[7] Elena Cabrio and Serena Villata. 2014. Towards a Benchmark of Natural Language Arguments. *CoRR* abs/1405.0 (2014).

[8] William Cai, Joshua Grossman, Zhiyuan Lin, Hao Sheng, Johnny Tian, Zheng Wei, Joseph Jay Williams, and Sharad Goel. 2019. MathBot: A Personalized Conversational Agent for Learning Math. (2019). https://doi.org/10.475/123{_}4

[9] Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. TARGER: Neural Argument Mining at Your Fingertips. (2019), 195–200. https://doi.org/10.18653/v1/p19-3031

[10] Michelene T.H. Chi and Ruth Wylie. 2014. The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist* 49, 4 (2014), 219–243. https://doi.org/10.1080/00461520.2014.965823

[11] 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.

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

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

[14] 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

[15] 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.

[16] 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

[17] Nicholas Diana. 2018. Leveraging educational technology to improve the quality of civil discourse. In *International Conference on Artificial Intelligence in Education*, Vol. 10948 LNAI. Springer Verlag, 517–520. https://doi.org/10.1007/978-3-319-93846-2{_}97

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

[19] 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.

[20] 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. https://doi.org/10.4324/9780203811306

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

[22] 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. https://doi.org/10.1080/00461520.2012.748005

[23] 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. https://doi.org/10.1024//0170-1789.20.4.309

[24] R Flesch. 1943. Marks of readable style; a study in adult education. *Teachers College Contributions to Education* 897 (1943).

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

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

[27] 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

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

[29] Sebastian Hobert. 2019. Say hello to 'Coding Tutor'! Design and evaluation of a chatbot-based learning system supporting students to learn to program. *40th International Conference on Information Systems, ICIS 2019* (2019), 1–17.

[30] Sebastian Hobert and Raphael Meyer Von Wolff. 2019. Say Hello to Your New Automated Tutor – A Structured Literature Review on Pedagogical Conversational Agents. *14th International Conference on Wirtschaftsinformatik, Siegen, Germany* (2019).

[31] 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. https://doi.org/10.1080/10494820.2015.1016533

[32] 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. https://doi.org/10.1007/s11423-009-9143-8

[33] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing data from chatbot and web surveys effects of platform and conversational style on survey response quality. *Conference on Human Factors in Computing Systems - Proceedings* (2019), 1–12. https://doi.org/10.1145/3290605.3300316

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

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

[36] Deanna Kuhn. 1993. Science as argument: Implications for teaching and learning scientific thinking. *Science Education* 77, 3 (6 1993), 319–337. https://doi.org/10.1002/sce.3730770306

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

[38] James A. Kulik and J. D. Fletcher. 2016. Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Review of Educational Research* 86, 1 (2016), 42–78. https://doi.org/10.3102/0034654315581420

[39] Severin Landolt, Thiemo Wambsganss, and S Matthias. 2021. A Taxonomy for Deep Learning in Natural Language Processing. In *Hawaii International Conference on System Sciences (HICSS)*.

[40] John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics* 45, 4 (2019), 765–818. https://doi.org/10.1162/COLIa00364

[41] Matthew K.O. Lee, Christy M.K. Cheung, and Zhaohui Chen. 2005. Acceptance of Internet-based learning medium: The role of extrinsic and intrinsic motivation. *Information and Management* 42, 8 (2005), 1095–1104. https://doi.org/10.1016/j.im.2003.10.007

[42] Marco Lippi and Paolo Torroni. 2016. MARGOT: A web server for argumentation mining. *Expert Systems with Applications* 65 (2016), 292–303. https://doi.org/10.1016/j.eswa.2016.08.050

[43] 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.

[44] 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. https://doi.org/10.1007/s11412-007-9025-1

[45] 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. https://doi.org/10.1002/tea.21316

[46] Sabine Payr. 2003. The virtual university's faculty: An overview of educational agents. *Applied Artificial Intelligence* 17, 1 (1 2003), 1–19. https://doi.org/10.1080/713827053

[47] Reinhard Pekrun and Elizabeth J. Stephens. 2012. Academic emotions. *APA educational psychology handbook, Vol 2: Individual differences and cultural and contextual factors.* (10 2012), 3–31. https://doi.org/10.1037/13274-001

[48] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics (ACL), 1532–1543. https://doi.org/10.3115/v1/d14-1162

[49] 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

[50] Eric Ries. 2011. The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses.

[51] Roman Rietsche and Matthias Söllner. 2019. Insights into Using IT-Based Peer Feedback to Practice the Students Providing Feedback Skill. *Proceedings of the 52nd Hawaii International Conference on System Sciences* (2019). https://doi.org/10.24251/hicss.2019.009

[52] Heinrich Roth. 1970. Pädagogische Psychologie des Lehrens und Lernens. https://issuu.com/audio2brain/docs/name6bce04

[53] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2019. QuizBot: A Dialogue-based Adaptive Learning System System for Factual Knowledge. Chi (2019), 1–13. https://doi.org/10.1145/3290605.3300587

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

[55] 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. https://doi.org/10.1007/s11412-009-9080-x

[56] Julia E. Seaman, I. E. Allen, and Jeff Seaman. 2018. *Higher Education Reports - Babson Survey Research Group*. Technical Report. http://www.onlinelearningsurvey.com/highered.htmlhttps://www.onlinelearningsurvey.com/highered.html

[57] Bayan Abu Shawar and Eric Steven Atwell. 2005. Using corpora in machine-learning chatbot systems. *International Journal of Corpus Linguistics* 10, 4 (2005), 489–516. https://doi.org/10.1075/ijcl.10.4.06sha

[58] Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*. 69–78.

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

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

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

[62] 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. https://doi.org/10.1007/s11251-011-9174-5

[63] Patrick Suppes and Mona Morningstar. 1969. Computer-assisted instruction. *Science* 166, 3903 (1969), 343–350. https://doi.org/10.1126/science.166.3903.343

[64] 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

[65] Heikki Topi. 2018. Using competencies for specifying outcome expectations for degree programs in computing: Lessons learned from other disciplines. *2018 SIGED International Conference on Information Systems Education and Research* (2018).

[66] Stephen E. Toulmin. 2003. *The uses of argument: Updated edition*. 1–247 pages. https://doi.org/10.1017/CBO9780511840005

[67] Viswanath Venkatesh and Hillol Bala. 2008. Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences* 39, 2 (5 2008), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

[68] 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.

[69] Jan vom Brocke, Alexander Simons, Kai Riemer, Bjoern 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), 205–224. https://doi.org/10.17705/1cais.03709

[70] Lev Semenovich Vygotsky. 1980. *Mind in society: The development of higher psychological processes*. Harvard university press.

[71] 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.

[72] Thiemo Wambsganss, Nikolaos Molyndris, and Matthias Söllner. 2020. Unlocking Transfer Learning in Argumentation Mining: A Domain-Independent Modelling Approach. In *15th International Conference on Wirtschaftsinformatik*. Potsdam, Germany. https://doi.org/10.30844/wi{_}2020{_}c9-wambsganss

[73] Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Jan Marco Leimeister, and Siegfried Handschuh. 2020. AL : An Adaptive Learning Support System for Argumentation Skills. In *ACM CHI Conference on Human Factors in Computing Systems*. 1–14.

[74] Thiemo Wambsganss, Christina Niklaus, Siegfried Handschuh, and Jan Marco Leimeister. 2020. Annotating Arguments and their Relations in Student Peer-Feedbacks. *Under review at COLING2020* (2020).

[75] Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. A Corpus for Argumentative Writing Support in German. In *28th International Conference on Computational Linguistics (Coling)*.

[76] Thiemo Wambsganss and Roman Rietsche. 2020. Towards designing an adaptive argumentation learning tool. In *40th International Conference on Information Systems, ICIS 2019*. Munich, Germany, 1–9.

[77] Thiemo Wambsganss, Florian Weber, and Matthias Söllner. 2021. Design and Evaluation of an Adaptive Empathy Learning Tool. In *Hawaii International Conference on System Sciences (HICSS)*.

[78] Thiemo Wambsganss, Rainer Winkler, Pascale Schmid, and Matthias Söllner. 2020. Designing a Conversational Agent as a Formative Course Evaluation Tool. In *15th International Conference on Wirtschaftsinformatik*. Potsdam, Germany.

[79] Thiemo Wambsganss, Rainer Winkler, Pascale Schmid, and Matthias Söllner. 2020. Unleashing the Potential of Conversational Agents for Course Evaluations: Empirical Insights from a Comparison with Web Surveys. In *Twenty-Eighth European Conference on Information Systems (ECIS2020)*. Marrakesh, Morocco, 1–18.

[80] Thiemo Wambsganss, Rainer Winkler, Matthias Söllner, and Jan Marco Leimeister. 2020. A Conversational Agent to Improve Response Quality in Course Evaluations. In *ACM CHI Conference on Human Factors in Computing Systems*.

[81] World Economic Forum WEF. 2018. *The Future of Jobs Report 2018*. Technical Report. https://doi.org/10.1177/0891242417690604

[82] 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. https://doi.org/10.1016/j.compedu.2005.04.003

[83] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent. In *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3313831.3376781

[84] R. Winkler and M. Söllner. 2018. Unleashing the Potential of Chatbots in Education : A State-Of-The-Art Analysis . In : Academy of Management. *Meeting, Annual Chicago, A O M* (2018). https://www.alexandria.unisg.ch/254848/1/JML_699.pdf

[85] N. Zierau, T Wambsganss, Andreas Janson, Sofia Schöbel, and Jan Marco Leimeister. 2020. The Anatomy of User Experience with Conversational Agents : A Taxonomy and Propositions of Service Clues. In *International Conference on Information Systems (ICIS)*. 1–17.