Design and Evaluation of an Adaptive Dialog-Based Tutoring System for Argumentation Skills

Thiemo Wambsganß
University of St.Gallen
29 PUBLICATIONS **87** CITATIONS
SEE PROFILE

Matthias Söllner
Universität Kassel
192 PUBLICATIONS **1,257** CITATIONS
SEE PROFILE

Jan Marco Leimeister
University of St.Gallen
965 PUBLICATIONS **10,061** CITATIONS
SEE PROFILE

Design and Evaluation of an Adaptive Dialog-Based Tutoring System for Argumentation Skills

Completed Research Paper

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

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

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 not only bear the opportunity to design new dialog-based forms of human-computer interaction but also to analyze the argumentation quality of texts. Both can be leveraged to provide students with individual and adaptive tutoring in their personal learning journey to develop argumentation skills. Therefore, we present the results of our design science research project on how to design an adaptive dialog-based tutoring system to help students to learn how to argue. Our results indicate the usefulness of an adaptive dialog-based tutoring system to support students individually, independent of a human instructor, time and place. In addition to providing our embedded software artifact, we document our evaluated design knowledge as a design theory. Thus, we provide the first step toward a nascent design theory for adaptive conversational tutoring systems to individual support metacognition skill education of students in traditional learning scenarios.

Keywords*:* dialog-based learning systems, pedagogical conversational agents, adaptive learning, argumentation learning, argumentation mining

Introduction

Today, information can constantly be accessed, so people need to develop skills other than the replication of information. Therefore, the requirements of job profiles are shifting towards more interdisciplinary, ambiguous and creative tasks (vom Brocke et al. 2015). As a result, educational institutions are called to evolve in their curricula when it comes to the compositions of skills and knowledge conveyed (Topi 2018). Most notably, teaching *metacognition skills* to students, such as critical thinking, collaboration or problem-solving, have become a central interested of educators (Fadel et al. 2015). The Organization for Economic Cooperation and Development (OECD) also recognized the importance of teaching these skills and thus included them as a major element of their Learning Framework 2030 (OECD 2018). One subclass of metacognition skills represents the skill of arguing in a structured, reflective and well-formed way (Toulmin 2003). 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 (Kuhn 1992). 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 strategic decision-making and analyzing different standpoints especially in regard to managing digitally enabled organizations. 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 (Black and Wiliam 2009; Hattie and Timperley 2007). 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 how to argue (Vygotsky 1980). 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, Seaman et al. 2018). One possible solution to imitate meaningful, individual *instructor–learner interactions* are *pedagogical conversational agents* (PCAs, e.g., Hobert and Wolff 2019), which have been successfully used in adaptively supporting learners to conduct a task by mimicking the gold standard of human tutors (e.g., Winkler et al. 2019). PCAs are software programs that communicate with users through natural language interaction interfaces (Shawar and Atwell 2005). By using an adaptive PCA as a tutor for argumentation skill learning, students would be able to conduct argumentative writing tasks and learn autonomously and independently of the instructor, time and place (Wambsganss and Rietsche 2020; Winkler and Söllner 2018).

Researchers, 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., De Groot et al. 2007; Osborne et al. 2016; Pinkwart et al. 2009). However, recent advances in Natural Language Processing (NLP) and Machine Learning (ML) to design and evaluate new forms of *human–computer interaction* (HCI) have not been assessed for argumentation skill learning. A new pedagogical HCI scenario could be to employ adaptive PCAs, to intelligently *tutor students in their individual argumentation learning process*, e.g., with adaptive and instant feedback, theoretical input, or a step-by-step guidance. In fact, the successful application of adaptive PCAs to meet individual needs of learners and to increase their learning outcomes has been demonstrated for learning various skills such as problem-solving skills (Winkler et al. 2019), programming skills (Hobert 2019), mathematical skills (Cai et al. 2019) as well as for learning factual knowledge (Ruan et al. 2019), and also offers potential for training argumentation skills. A possible solution to provide adaptive support for argumentation learning could be the utilization of argumentation mining (AM), an approach from the field of *Computational Linguistics* to identify and classify argumentation in texts, e.g., to access individual levels of argumentation (Wambsganss, Niklaus, Cetto, et al. 2020; Wambsganss and Rietsche 2020). The potential of AM has been investigated in different research domains, such as accessing argumentation flows in legal texts (Moens et al. 2007), getting a deeper understanding of customer opinions in user-generated comments (Boltuži and Šnajder 2014) or fact-checking and de-opinionizing of news (Dusmanu et al. 2018). However, AM has not been used for accessing the skill level of students to provide adaptive and individual tutoring through a dialog-based interaction with a PCA, such as providing adaptive feedback on the persuasiveness of texts or guiding a learner through writing an argumentative text (Lawrence and Reed 2019). Leveraging the technological advances of AM, NLP and ML, we aim to address this gap by designing and evaluating a new form of *human–computer interaction* for argumentation skill learning. We aim to generate design knowledge for an *adaptive dialog-based argumentation learning tutoring system* that assists students on an individualized level in writing argumentative texts and supports the development of their argumentation skills. Our goal is to provide students with assistance comparable to a discussion with a human tutor. Overall, we aim to contribute by answering the following research questions (RQ):

RQ1: *What requirements should be considered when designing an adaptive dialog-based argumentation tutoring system that aims to support students in writing argumentative texts?*

RQ2: *How useful is an adaptive dialog-based argumentation tutoring system for students to learn how to argue?*

To answer the stated research questions, we follow the *design science research approach* (DSR) by Hevner (2007). We intend to iteratively design and evaluate an *adaptive dialog-based tutoring system* that aims to support students in writing argumentative texts through individual argumentation tutoring and feedback informing the artifact design (Hevner et al. 2004). We refer to the terms adaptive PCA and adaptive dialog-based tutoring system as synonyms in our paper. The remainder is structured as follows: In the next section, we explain the theoretical background. In our research endeavor, we build on the *ICAP framework* as our guiding kernel theory (Chi and Wylie 2014). Afterwards, we outline our research design following DSR and describe in detail how we design our adaptive dialog-based tutoring system *ArgueTutor* based on both scientific literature as well as insights from the field. By describing our design and evaluation process in eight consecutive steps, we document the generated design knowledge as a *nascent design theory* as proposed by Gregor and Hevner (2013) and Jones and Gregor (2007). Finally, we summarize and discuss our results, limitations and future research.

Theoretical Background

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 (Eemeren et al. 1996). Logical, structured arguments are a required precondition for persuasive conversations, general decision-making and drawing acknowledged conclusions. In the context of digitalization, the ability to argue becomes increasingly important for successful collaboration in almost every job, since job profiles are shifting towards interdisciplinary, ambiguous and creative tasks (vom Brocke et al. 2018). This has been recognized by the OECD, which named these metacognition skills a major part of their Learning Framework 2030 (OECD 2018). Not only in industry but also in research, studies show that argumentation is central to scientific thinking (Duschl and Osborne 2002; Kuhn 1993). As von Aufschnaiter *et al.* (2008) describes, scientists engage in argumentation to articulate, refine and discuss their scientific statements and the ones of others. According to Kuhn (1992), 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: *"It is in argument that we are likely to find the most significant way in which higher order thinking and reasoning figure in the lives of most people. Thinking as argument is implicated in all of the beliefs people hold, the judgments they make, and the conclusions they come to; it arises every time a significant decision must be made. Hence, argumentative thinking lies at the heart of what we should be concerned about in examining how, and how well, people think*" (Kuhn 1992, pp. 156–157).

However, the integration of elements that aim at teaching argumentation in learning scenarios is limited. Jonassen and Kim (2010), p. 442 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 (Driver et al. 2000; Kuhn 2005; Scheuer et al. 2012; Stab and Gurevych 2017; Wambsganss and Rietsche 2020). Most students simply learn to argue in the course of their studies 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 (Hattie and Timperley 2007; Vygotsky 1980). 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 (Scheuer 2015). Especially in increasingly common large-scale lectures or distance learning settings such as MOOCs (Seaman et al. 2018), 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 and feedback for a single student. As a consequence, many researchers have designed and evaluated several technology-enhanced learning tools based on input masks and representational guidelines to support the active writing process of high school students. This has been investigated across a variety of fields, including law (Pinkwart et al. 2009), science (Osborne et al. 2016; Suthers and Hundhausen 2001) and conversational argumentation (De Groot et al. 2007). However, recent advances in NLP, ML or AM to design and 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, has merely been assessed for adaptive argumentation skill learning (Stab and Gurevych 2017; Wambsganss and Rietsche 2020). 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 (Hobert 2019; Wambsganss and Rietsche 2020).

Argumentation Mining for Adaptive Argumentation 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 (Toulmin 2003). Toulmin's model asserts that a "*good*" argument involves a logical structure built on *ground*, *claim* and *warrant*, whereas the *grounds* are the evidence used to prove a *claim*. Walton et al. (2008) developed the so-called *"argumentation schemes"* that use Toulmin's type of reasoning. It is commonly considered that "*Claim*", "*Premise*" and "*Warrant*" are the main components of every argument, the rest are supporting sub-argument parts that can exist in an argument. AM, a research field in Computational Linguistics, aims at automatically identifying arguments in unstructured texts (Lawrence and Reed 2019). In the identification of these argumentation structures, two main tasks can be distinguished:

- **Argument component classification,** classification of argumentative text into claims and premises, and
- **Argumentative discourse analysis,** identification of *support* and *attack* relationships between pairs of argument components.

The potential of AM has been investigated in different research domains, such as accessing argumentation flows in legal texts (Moens et al. 2007), better understanding of customer opinions in user-generated comments (Boltuži and Šnajder 2014), or fact-checking and de-opinionizing of news (Dusmanu et al. 2018). Researchers have built increasing interest in intelligent writing assistance (Song et al. 2014; Stab and Gurevych 2014, 2017), since it enables adaptive argumentative writing support with tailored feedback about arguments in texts (Wambsganss, Niklaus, Cetto, et al. 2020; Wambsganss and Rietsche 2020). However, the complexity of using this technology in a dialog-based learning scenario for educational purposes has poorly been assessed yet (Lawrence and Reed 2019).

Pedagogical Conversational Agents for Adaptive Argumentation Learning

Recent advances in NLP and ML bear the opportunity to design new forms of *human–computer interaction* with conversational interfaces, also called Conversational Agents (CA). CAs are software programs that are designed to communicate with users through natural language interaction interfaces (Shawar and Atwell 2005). In today's world, conversational interfaces, such as *Amazon's Alexa*, *Google's Assistant* or *Apple's Siri*, are ubiquitous, with their popularity steadily growing over the past few years (Krassmann et al. 2018). They are implemented in various areas, such as customer service (Hu et al. 2018), counselling (Cameron et al. 2017) or education (e.g., Wambsganss, Winkler, et al. 2020a). Hobert and Wolff (2019) define CAs used in education as a special form of learning application that interacts with learners individually. 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) (e.g., Atkinson and Shiffrin 1968; Suppes and Morningstar 1969). Similar to a human tutor, these systems can present instructions, ask questions and provide immediate feedback (Kulik and Fletcher 2016). 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 (Payr 2003), motivators or learning companions (Hobert and Wolff 2019) as well as conducting course evaluations to assist teachers (Wambsganss et al. 2020; Wambsganss, Winkler, Schmid, et al. 2020b). 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 (Winkler et al. 2019), programming skills (Hobert 2019), mathematical skills (Cai et al. 2019) as well as for learning factual knowledge (Ruan et al. 2019), but it has not been investigated for argumentation skills. An adaptive PCA could offer new forms of providing individual argumentative guidance and feedback to students, e.g., when completing a task to write persuasive texts, through a natural conversation interface combined with AM technology. However, in *literature exists no approach with principles, design knowledge and evaluation on how to design and embed an adaptive dialog-based argumentation tutoring system in a pedagogical scenario to help students to learn how to argue* (Hobert and Wolff 2019; Winkler and Söllner 2018).

ICAP Framework as a Kernel Theory to Interactively Engage Students in Learning

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 Chi and Wylie (2014), the learners' engagement with learning materials can range *"from passive to active to constructive to interactive"* (Chi and Wylie 2014) 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 Chi and Wylie (2014), 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 (Chi and Wylie 2014). 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. The application of the ICAP Framework has been successfully demonstrated to interactively engage students in learning problem-solving skills (Winkler et al. 2019) or programming skills (Hobert 2019). Therefore, we believe that a user-centered and literature-based design of an adaptive PCA combined with intelligent algorithms to provide individual learning tutoring for argumentation skills by individually assisting students in writing persuasive texts would interactively engage them according to the ICAP Framework.

Research Methodology

Our study is guided by the DSR approach (Hevner 2007). We decided to follow this approach in order to use a scientific method to solve a set of practical problems that researchers and practitioners experience in their own practice and to contribute to the existing body of knowledge by designing and evaluating a new research artifact. Figure 1 shows the eight consecutive steps that have been conducted to ensure design knowledge is created on insights from the application domain and the knowledge base. We analyze requirements based on the environment (*relevance*) and knowledge base (*rigor*), derive design principles, and instantiate and evaluate them with our adaptive PCA *ArgueTutor* in two design cycles. Our project contributes to research with a *nascent design theory* that gives explicit prescriptions for designing this class of artifacts (Gregor and Hevner 2013). We followed a theory-driven design approach by grounding our research on the ICAP framework by Chi and Wylie (2014). The ICAP framework therefore motivates the overall design and evaluation of our DSR approach. The *first step* of the DSR cycle includes the problem formulation. The relevance of the practical problem was therefore described in the introduction and the theoretical background of this work. In the *second step*, we derived a set of meta-requirements (MRs) from the current state of scientific literature for the design of adaptive PCAs for argumentation learning. Next, we conducted twelve semi-structured interviews with students, using the expert interview method by Gläser and Laudel (2010). Based on the interviews, we gathered user stories (USs) and user requirements (URs) for the design of an adaptive PCA for argumentation learning following (Cohn 2004). In the *fourth step*, we derived design principles (DPs) addressing the MRs and URs from the prior steps, using the structure suggested by Gregor et al. (2020) and designed an initial version of our PCA called *ArgueTutor* with design features as a first instantiation of these DPs. In the *fifth step*, we followed the evaluation framework proposed by Venable et al. (2016) and performed a *formative* and *artificial* evaluation of our initial version with 32 students to collect feedback in the early design stage of our dialog-based tutoring system. Based on these findings, we refined our design knowledge, such as the instantiated design features, added an additional design principle and built a fully functional software artifact as a second version of *ArgueTutor,* which supports and assists student on an individualized level in writing argumentative texts through guidance and adaptive argumentation feedback independent of an educator, time and location. We designed a conversational interface with meaningful interaction dialogs based on our DPs and a fully functional back end algorithm based on our argumentation-annotated corpora of student-written essays and an AM model that provides adaptive and instant feedback on the level of argumentation and the persuasiveness of a submitted student text. In the *seventh step*, we evaluate our fully working version of

ArgueTutor following a *naturalistic ex post* evaluation by Venable et al. (2016) to answer our second research question (**RQ2**): *How can an adaptive dialog-based argumentation tutoring system for students be designed to learn how to argue?* Thus, we implement two different versions of *ArgueTutor* (one providing individual argumentation support one, one providing nonadaptive argumentation support) in an experiment where 45 students had to write a persuasive text and received tutoring by *ArgueTutor*.

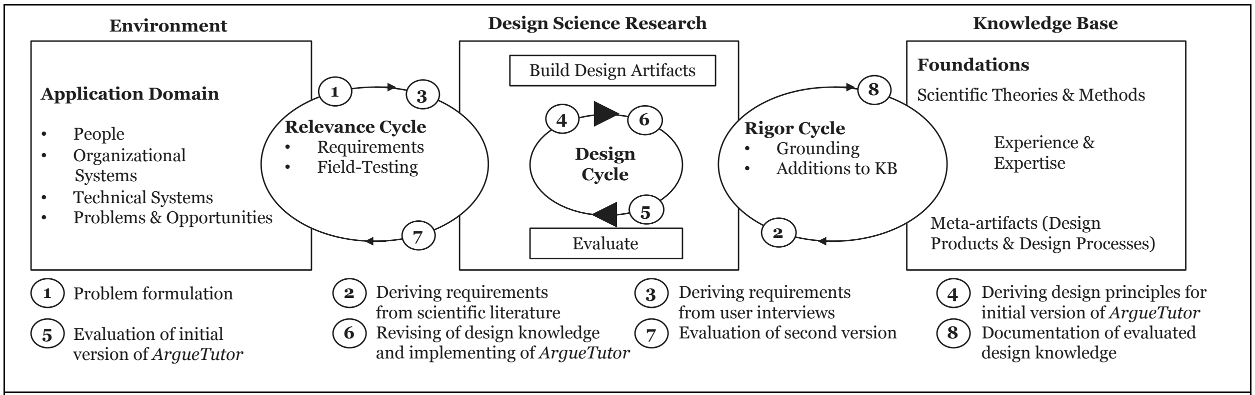

Environment
Design Science Research
Knowledge Base
Application Domain
• People
• Organizational Systems
• Technical Systems
• Problems & Opportunities
1
3
Relevance Cycle
• Requirements
• Field-Testing
7
Build Design Artifacts
4
6
Design Cycle
5
Evaluate
8
Rigor Cycle
• Grounding
• Additions to KB
2
Foundations
Scientific Theories & Methods
Experience & Expertise
Meta-artifacts (Design Products & Design Processes)
1 Problem formulation
5 Evaluation of initial version of ArgueTutor
2 Deriving requirements from scientific literature
6 Revising of design knowledge and implementing of ArgueTutor
3 Deriving requirements from user interviews
7 Evaluation of second version
4 Deriving design principles for initial version of ArgueTutor
8 Documentation of evaluated design knowledge

Figure 1. Three cycle design science process according to Hevner (2007)

Therefore, we contribute to research with an evaluated learning tool that can be used in a pedagogical scenario where students finish a certain exercise in a lecture (e.g., writing a persuasive peer review to a fellow student) and additionally receive adaptive tutoring based on their argumentation level through a PCA. The evaluated design knowledge from this research project is summarized as a *nascent design theory* (Gregor and Hevner 2013) for dialog-based learning applications to support argumentation skill learning.

Design and Evaluation of *ArgueTutor*

In this section, we will describe and discuss how we designed and evaluated our adaptive PCA *ArgueTutor* in eight consecutive steps following DSR. The problem formulation (*step one*), described in the *introduction* and in the *theoretical background*, serves as the foundation for the derivation of the requirements.

Step 2 & 3: Deriving Requirements from Scientific Literature and User Interviews

To derive requirements from scientific literature, a systematic literature search was conducted using the methodological approaches of Cooper (1988) and vom Brocke et al. (2015). Based on that, we (*1*) defined the review scope, (*2*) conceptualized the topic, (*3*) searched the literature and (*4*) analyzed the findings regarding requirements. Regarding step *1*, defining the review scope, we primarily focused our research on studies that demonstrate the successful implementation of PCAs and argumentation skill learning. Furthermore, our goal is to identify requirements on a conceptual level with a focus on an espousal of position and a representative coverage (Cooper 1988). Regarding step *2,* conceptualizing the topic, we identified two broad areas for deriving requirements: *Educational Technology* and *Human–Computer Interaction*. Since the creation of an adaptive PCA 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. Regarding step *3,* literature search, we conducted a keyword search on Google Scholar to identify relevant publications. We used Google Scholar because this web search engine enables advanced full-text search and several filter options for academic literature. We used the following search terms: *„Argumentation Learning”, “Reasoning Skills”, “Pedagogical Conversational Agent”, „Skill Learning”, „Learning Theory“, „Learning Tutor”, “Dialog-based Learning*” and “*Learning Theories*”. We defined criteria for inclusion and exclusion and reviewed titles and abstracts of our search results in a first step. We only included literature that deals with or contributes to a kind of dialog-based learning tool in the field of argumentation learning, such as an established learning theory. Several papers that have dealt with argumentation or conversational agents in other research areas than education were excluded. On this basis, we selected 85 papers for more intensive analysis. We have summarized similar topics of these contributions as *literature issues* (LIs) and formed five clusters from them (e.g., LI1 (Soloway et al. 1996) or LI2 (Nass and Moon 2000)). Based on these LIs, we derived *meta-requirements* (MRs) for the design of a PCA for argumentation learning. The resulting clusters are illustrated as literature issues in Figure 2. Based on the derived LIs and MRs, we conducted twelve semi-structured interviews according to Gläser and Laudel (2010). 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 PCA for argumentation skills. The participants were asked about the following topics: *experience with technology-based learning systems*, *perception of existing learning systems in use*, *experiences with conversational agents, requirements for an adaptive PCA that supports writing argumentative texts* (e.g., functionalities, design), *requirements for an adaptive PCA that supports learning how to argue through individual writing support* (e.g., functionalities, design).

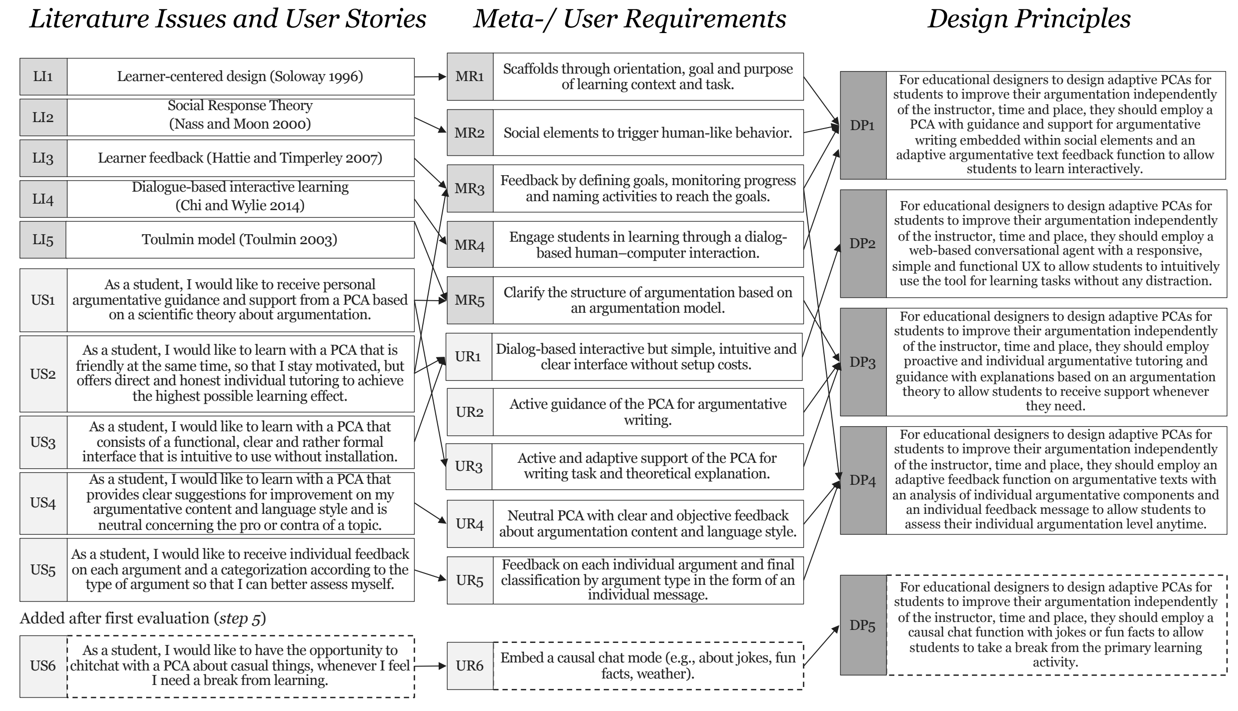

Literature Issues and User Stories
Meta-/ User Requirements
Design Principles
LI1 Learner-centered design (Soloway 1996)
LI2 Social Response Theory (Nass and Moon 2000)
LI3 Learner feedback (Hattie and Timperley 2007)
LI4 Dialogue-based interactive learning (Chi and Wylie 2014)
LI5 Toulmin model (Toulmin 2003)
US1 As a student, I would like to receive personal argumentative guidance and support from a PCA based on a scientific theory about argumentation.
US2 As a student, I would like to learn with a PCA that is friendly at the same time, so that I stay motivated, but offers direct and honest individual tutoring to achieve the highest possible learning effect.
US3 As a student, I would like to learn with a PCA that consists of a functional, clear and rather formal interface that is intuitive to use without installation.
US4 As a student, I would like to learn with a PCA that provides clear suggestions for improvement on my argumentative content and language style and is neutral concerning the pro or contra of a topic.
US5 As a student, I would like to receive individual feedback on each argument and a categorization according to the type of argument so that I can better assess myself.
Added after first evaluation (step 5)
US6 As a student, I would like to have the opportunity to chitchat with a PCA about casual things, whenever I feel I need a break from learning.
MR1 Scaffolds through orientation, goal and purpose of learning context and task.
MR2 Social elements to trigger human-like behavior.
MR3 Feedback by defining goals, monitoring progress and naming activities to reach the goals.
MR4 Engage students in learning through a dialog-based human–computer interaction.
MR5 Clarify the structure of argumentation based on an argumentation model.
UR1 Dialog-based interactive but simple, intuitive and clear interface without setup costs.
UR2 Active guidance of the PCA for argumentative writing.
UR3 Active and adaptive support of the PCA for writing task and theoretical explanation.
UR4 Neutral PCA with clear and objective feedback about argumentation content and language style.
UR5 Feedback on each individual argument and final classification by argument type in the form of an individual message.
UR6 Embed a causal chat mode (e.g., about jokes, fun facts, weather).
DP1 For educational designers to design adaptive PCAs for students to improve their argumentation independently of the instructor, time and place, they should employ a PCA with guidance and support for argumentative writing embedded within social elements and an adaptive argumentative text feedback function to allow students to learn interactively.
DP2 For educational designers to design adaptive PCAs for students to improve their argumentation independently of the instructor, time and place, they should 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.
DP3 For educational designers to design adaptive PCAs for students to improve their argumentation independently of the instructor, time and place, they should employ proactive and individual argumentative tutoring and guidance with explanations based on an argumentation theory to allow students to receive support whenever they need.
DP4 For educational designers to design adaptive PCAs for students to improve their argumentation independently of the instructor, time and place, they should employ an adaptive feedback function on argumentative texts with an analysis of individual argumentative components and an individual feedback message to allow students to assess their individual argumentation level anytime.
DP5 For educational designers to design adaptive PCAs for students to improve their argumentation independently of the instructor, time and place, they should employ a causal chat function with jokes or fun facts to allow students to take a break from the primary learning activity.

Figure 2. Overview of the derived design principles according to Gregor et al. (2020).

The interviewed students were between 21 and 27 years old; nine were enrolled in a master program, three in a bachelor program at our university. All participants were business students; 6 were male, 6 were female. The interviews were coded and abstract categories were formed. The coding was performed using open coding to form a uniform coding system during evaluation (Gläser and Laudel 2010). Based on these results, we gathered *user stories* (USs), aggregated the most common ones and identified *user requirements* (URs) following Cohn (2004). Our results are illustrated in Figure 2.

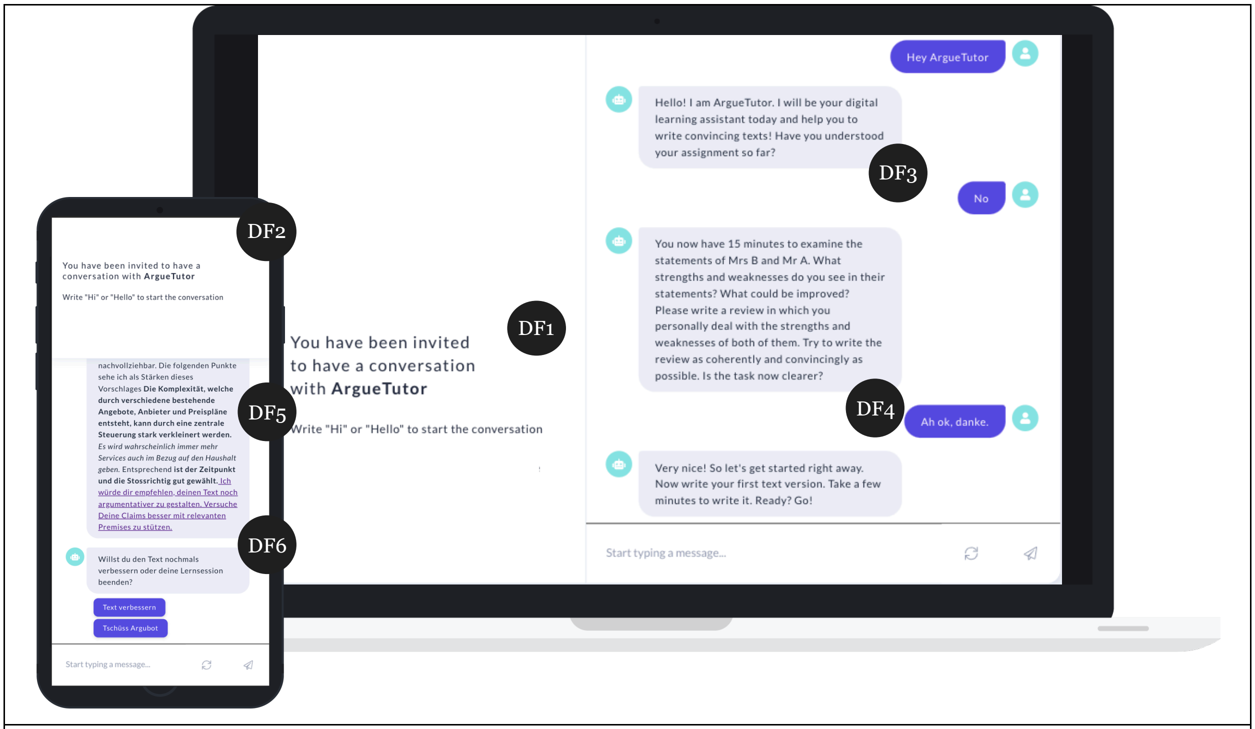
Step 4: Deriving Design Principles and Design Features for ArgueTutor

As illustrated, we have identified a concise set of LIs, USs and formulated MRs and URs. Based on these findings, we derived four preliminary *design principles* (DPs) for an adaptive PCA for argumentation learning, a special class of adaptive dialog-based learning tools for metacognition skills. The DPs are depicted in Figure 2. For formulating the DPs, we relied on the conceptual schema of Gregor et al. (2020) and therefore believe they are self-explanatory. However, we will discuss them further in the evaluation.

To instantiate and evaluate our DP, we created a first prototype of *ArgueTutor*[1] with *design features* (DF) as instantiations of our DPs. In the first version, we implemented a clickable mock-up of *ArgueTutor* to receive initial feedback in an early design stage on the overall concept. The mock-up illustrated individual

[1] *ArgueTutor* was designed in German to provide German students with adaptive guidance and feedback. However, for ease of understanding in this paper, we translated parts of the dialogs into English.

tutoring of a student by a PCA to support the writing of argumentative texts. The fully functional software artifact was implemented in *Step 6*. It consists of a user-centered front-end and a back-end algorithm based on rule-based conversational intents and an AM mining model to provide adaptive feedback and individual guidance based on the argumentation skill level of students. Based on **DP1** *and* **DP2**, we only included simple design elements in the design of *ArgueTutor (***DF2***)*, since we aimed to offer a more formal and functional tutoring experience based on **US3** that does not distract students. However, we believe this DP 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*.


You have been invited to have a conversation with **ArgueTutor**

Write "Hi" or "Hello" to start the conversation

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 of them. Try to write the review as coherently and convincingly as possible. Is the task now clearer?

Ah ok, danke.

Very nice! So let's get started right away. Now write your first text version. Take a few minutes to write it. Ready? Go!

You have been invited to have a conversation with **ArgueTutor**

Write "Hi" or "Hello" to start the conversation

nachvollziehbar. Die folgenden Punkte sehe ich als Stärken dieses Vorschlages **Die Komplexität, welche durch verschiedene bestehende Angebote, Anbieter und Preispläne entsteht, kann durch eine zentrale Steuerung stark verkleinert werden.** *Es wird wahrscheinlich immer mehr Services auch im Bezug auf den Haushalt geben.* Entsprechend **ist der Zeitpunkt und die Stossrichtig gut gewählt.** Ich würde dir empfehlen, deinen Text noch argumentativer zu gestalten. Versuche Deine Claims besser mit relevanten Premises zu stützen.

Willst du den Text nochmals verbessern oder deine Lernsession beenden?

Figure 3. Screenshot of the interface of *ArgueTutor* with exemplars of design features

We built *ArgueTutor* as a responsive web-based application that can be used on all kinds of devices to instantiate **DP2**. The front end of *ArgueTutor* was developed with recent web technologies including HTML5, Cascading Style Sheets (CSS) and JavaScript (JS). Students are able to access *ArgueTutor* from any web-enabled device regardless of its screen size or operation system (**DF1**). For implementing the conversational logic of *ArgueTutor, we* relied on the chatbot framework *Rasa* (Bocklisch et al. 2017)*,* since it is open source and supports major NLP language models to implement intelligent chat intents. To instantiate **DP3 and DP4,** we design *ArgueTutor* with an adaptive argumentation learning mode (**DF3 – DF6**). *ArgueTutor* guides students through writing a persuasive text with the aim to imitate a human educator (**DF3 and DF4**). The PCA proactively explains a writing task (e.g., students have to write persuasive peer feedback to a fellow student) (**DF3**) and provides hints and explanations when the user asks for help such as argumentation theory input (Toulmin 2003) (**DF4**). 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 (**DF5**) and by tutoring students with an individual feedback message on what to improve depending on the student's skill level (**DF6**). The adaptivity of the argumentation feedback is implemented by training a supervised machine learning model based on our corpora of persuasive student-written texts (further explained in *step 6*). 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* (Bocklisch et al. 2017). A screenshot of *ArgueTutor* with exemplary instantiated design features (e.g., DF1 – DF7) can be seen in Figure 3 and 4.

Step 5: Evaluation of Initial Version of ArgueTutor

In this section, we describe the evaluation of the initial version of *ArgueTutor*. Based on the gathered requirements and design principles, we designed a clickable mock-up displaying a conversational interface with a step-by-step guidance for writing argumentative texts and individual argumentation feedback (without implementing trained chat intents in the back end). For the evaluation, we followed an *ex ante* evaluation using an *artificial* evaluation setup as proposed by Venable et al. (2016). The purpose of the evaluation is to check whether the design principles are useful for learners, in order to incorporate any change requests. The design principles were specifically examined based on the criteria of usefulness and usability and evaluated by means of various questions. Therefore, a questionnaire with 38 items was created with specific questions about the DP and the DF, e.g., we asked questions about the perception of the argumentation theory input, the usability and usefulness of the adaptive tutoring functions and the concept of the adaptive PCA as a whole. The participants received a link with a clickable mock-up of *ArgueTutor* as a web app. After testing the concept of our PCA, the participants were asked to complete a survey. In total, 32 participants successfully tested our prototypical version of *ArgueTutor* and fully answered our survey. 22 participants were master students, eight were enrolled in a bachelor program, 2 had a different occupation. All were potential users of an adaptive dialog-based argumentation tutoring system. Their age ranged from 22 and 34 years; 19 were female, 13 male. In order to obtain detailed feedback on our DF and DP, the respondents were able to classify the usefulness of the DF and DP based on a 7-point Likert Scale, e.g., for usability, ranging from *"completely useful (1)"* to *"completely useless (7)"*. The evaluation of the general usability as well as usefulness yielded the following result: The respondents rated the implemented DF and the DP positively compared to the neutral value of 4 (mean of 2.45 of all DF and DP related questions). The overall concept of an adaptive PCA for argumentation learning was also rated as very useful (mean of 2.16, n=32). Also, the concept of the specific DPs each got very positive ratings. DF5 and DF6, for example, were rated as completely useful by 96.9% of the participants. The argumentative guidance and theoretical argumentation input (DF3 and DF4) were rated by 93.5% as very useful. Moreover, we asked the participants for some qualitative feedback and additional ideas to improve the concept of *ArgueTutor*. Here, a majority of users mentioned that they would like to have a casual chat function, e.g., that the PCA tells a joke or answer to *"how are you*?" to receive further motivation while learning. Based on the feedback, we created a sixth user story, resulting in **UR6** and **DP5** (see Figure 2). Moreover, almost all users mentioned that the accuracy of the adaptive argumentation feedback is very important for revising one's own argumentation quality of a text and thus to increase their argumentation.

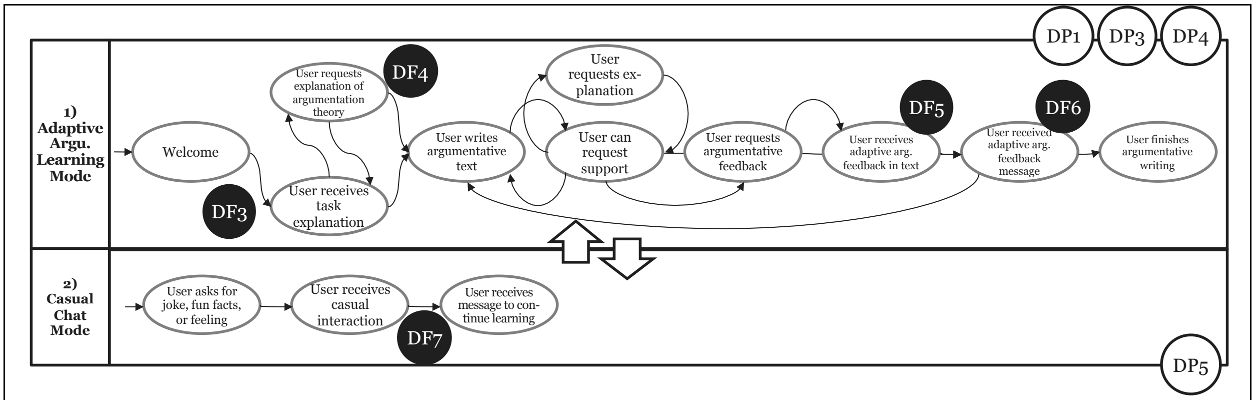

DP1
DP3
DP4
1) Adaptive Argu. Learning Mode
Welcome
DF3
User receives task explanation
User requests explanation of argumentation theory
DF4
User writes argumentative text
User requests ex-planation
User can request support
User requests argumentative feedback
DF5
User receives adaptive arg. feedback in text
DF6
User received adaptive arg. feedback message
User finishes argumentative writing
2) Casual Chat Mode
User asks for joke, fun facts, or feeling
User receives casual interaction
DF7
User receives message to con-tinue learning
DP5

Figure 4. Dialog path from a user perspective with exemplary design features

Step 6: Revising Design Knowledge and Implementing ArgueTutor

Following the feedback of the evaluation of our first version, we generated a new DP to include a casual chat function in our design. This DP is also backed with recent literature on adaptive PCAs where casual chat functions have been successfully implemented, e.g., Ruan et al. (2019). The final version of our PCA *ArgueTutor* consists of two interaction modes: 1) an *adaptive argumentation learning mode* that tutors students through support in argumentative writing and individual argumentation feedback, and 2) a *casual chat mode* (**DF7**). Moreover, we revised small details of almost all instantiated DF. Based on the revised version, our objective was to build and implement a fully functional software artifact. As a first step, we implemented the new design feature **DF7** to instantiate **DP5.** To do so, *ArgueTutor* incorporates a user-initiated and rule-based casual chat mode. A student can ask *ArgueTutor* to tell jokes or fun facts to take a break from the primary learning activity (see Figure 4 and 5) and thus change from the *argumentation learning mode* to the *casual chat mode* at any time she wants. To imitate the student 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 4.

Since we received much feedback that the accuracy and quality of the adaptive feedback on the persuasiveness of a message would be a crucial feature to use an adaptive PCA for argumentation learning, we rigorously focused on developing a state-of-the-art AM model that identifies argumentative structures in student-written texts in the forms of claims, premises and non-argumentative texts (**DP3** and **DP4**). Thus, we collected and annotated our own corpus to train and tune a model that fulfils the users' requirements to give individual and instant feedback on their texts and implemented it into *ArgueTutor*.

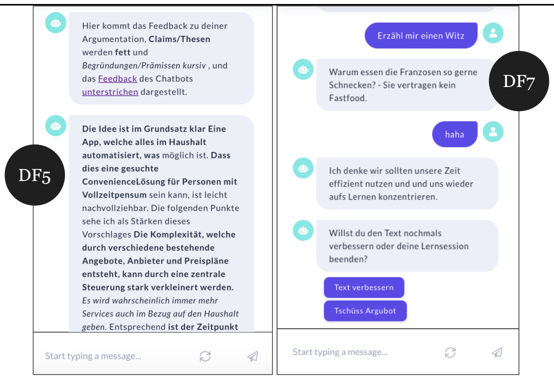
Building a Corpus of Argumentation-Annotated Student-Written Texts

A major prerequisite for developing supervised ML models based on NLP 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, which we firstly published and described in Wambsganss, Niklaus, Cetto, et al. 2020. The data was collected in one of our mandatory business innovation lectures in a master program at our university. In this lecture, around 220 students develop and present a new business model for which they receive three peer reviews each, and then a student from the same course elaborates on the strengths and weaknesses of the 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 from the last years. In the annotation process, we followed the approach described in Stab and Gurevych (2017a). 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 Stab and Gurevych (2017a), 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 (Stenetorp et al. 2012). After the first 100 reviews were annotated by all three annotators, we calculated the inter-annotator agreement scores. As we obtained satisfying results, we proceeded with a single annotator who marked up the remaining 900 documents. The joint unitized measure for claims and premises is α_U = 0.4096, suggesting a moderate agreement between the annotators.

Building the Adaptive Argumentation Mining Algorithm in the Back End

Guided by literature about AM and Text Mining, e.g., Fromm et al. 2019 and Wambsganss, Molyndris, et al. 2020, we built a predictive model following the architecture of Bidirectional Encoder Representations from Transformers (BERT) proposed by Devlin et al. (2018) to classify text tokens as a *claim*, *premise* or as *non-argumentative* following the argumentation theory of Toulmin (2003). We used the BERT model from *deepset*, 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 the FARM framework from *deepset*. 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 (**DP3** and **DP4**). We split the data into 70% training, 20% validation and 10% test data. The best performing model configuration is tested and evaluated on previously unseen test data (Bird et al. 2009). For the proposed architecture, the inputs and outputs 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 *warmup* and *embedding dropout probability* of *0.1* and *0.15* respectively. After several iteration, our final BERT model reached a macro f1 score of *0.73* percent for classifying text tokens into claim, premise or non-argumentative tokens. Compared to other studies on student-written argumentation identification (e.g., Eger et al. 2017; Stab and Gurevych 2017), this is a satisfying result.



Hier kommt das Feedback zu deiner Argumentation, **Claims/Thesen** werden **fett** und *Begründungen/Prämissen kursiv* , und das Feedback des Chatbots unterstrichen dargestellt.

Die Idee ist im Grundsatz klar Eine App, welche alles im Haushalt automatisiert, was möglich ist. **Dass dies eine gesuchte ConvenienceLösung für Personen mit Vollzeitpensum** sein kann, ist leicht nachvollziehbar. Die folgenden Punkte sehe ich als Stärken dieses Vorschlages **Die Komplexität, welche durch verschiedene bestehende Angebote, Anbieter und Preispläne entsteht, kann durch eine zentrale Steuerung stark verkleinert werden.** *Es wird wahrscheinlich immer mehr Services auch im Bezug auf den Haushalt geben*. Entsprechend **ist der Zeitpunkt**

Erzähl mir einen Witz

Warum essen die Franzosen so gerne Schnecken? - Sie vertragen kein Fastfood.

haha

Ich denke wir sollten unsere Zeit effizient nutzen und und uns wieder aufs Lernen konzentrieren.

Willst du den Text nochmals verbessern oder deine Lernsession beenden?

Figure 5. Adaptive learning mode (DF5) and casual chat mode (DF7) of *ArgueTutor*

To implement the adaptive argumentation tutoring (**DF5** and **DF6**), the model was exported and implemented in a back end chat intent of *ArgueTutor*. For this, a student enters a persuasive text and sends it to *ArgueTutor*, e.g., a potential argumentative text for a certain exercise (see Figure 4). The message is then sent to our trained model. *Claims*, *premises* and *non-argumentative* tokens are being classified and sent back to the front end. *ArgueTutor* provides an individual feedback by highlighting the argumentative components (e.g., claims and premises) of the texts and returning it to the student (**DF5**). Following Lippi and Torroni (2016a), claims are displayed in bold font, whereas premises are displayed in italic style (see Figure 5). Non-argumentative text paragraphs are not highlighted. Additionally, *ArgueTutor* provides an individual summarizing feedback based on the number of premises and claims in the message (**DF6** in Figure 3). 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. Both adaptive chat functions are implemented as chat intents in the back end of *ArgueTutor* following the *Rasa* framework for conversational interfaces (Bocklisch et al. 2017).

Step 7: Evaluation of Second Version of ArgueTutor

In *step 7,* we evaluated the second fully functional version of *ArgueTutor* in an online experiment to answer our second research question (**RQ2**): *How useful is an adaptive dialog-based argumentation tutoring system for students to learn how to argue?* Accordingly, we performed a *naturalistic ex post* evaluation of our second version of *ArgueTutor* (Venable et al. 2016). To achieve our goal, we aimed to evaluate 1) the overall usefulness of a dialog-based argumentation tutoring system and 2) the usefulness of the adaptive tutoring of *ArgueTutor*. To do so, we designed an experiment in which participants were asked to write a persuasive peer review based on a given essay. Participants were randomly assigned to two different groups in which we manipulated the level of adaptivity of our PCA. The treatment group (TG) used our fully functional *ArgueTutor* (with adaptive argumentation feedback on texts, *including* **DF5** and **DF6**), while participants in the control group (CG) used a version of *ArgueTutor* that only provides general feedback on texts (without our trained AM feedback intend – *not including* **DF5** and **DF6**). Instead participants of the

CG received a random general argumentation feedback based on the argumentation model of Toulmin (2003), e.g., *"please make sure all claims are sufficiently supported by at least one premise"*. We recruited 45 students from our university through social networks and mailing lists to take part in our experiment. After randomization, we had 22 participants in the treatment and 23 in the control group. The participants received a link to the experiment, which was conducted in the browser using the tool *unipark*. Participants of the treatment group had an average age of *24.85* (SD= *2.32*), 17 were male, 5 were female. In the control group, participants' average age was *25.39* (SD= *0.98*), 14 were male, 9 were female. In average the experiment took 25 to 35 minutes. It consisted of three main phases: *1) pretest phase, 2) individual writing phase* and 3) *post test phase*. The *pre-* and *post-phases* were consistent for all participants. In the *writing phase*, we manipulated the level of argumentation feedback participants received during writing their peer reviews.

Pretest 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 Agarwal and Karahanna (2000). Second, we tested the construct of *feedback-seeking of individuals* following Ashford (1986). 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 Flender et al. (1999), 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 *perceived argumentation learning*. 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 Flender et al. (1999). 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?"* (Flender et al. 1999). 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 Flender et al. (1999).

Individual 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 the remaining time. They were only able to continue the experiment after the countdown was finished. The treatment group used *ArgueTutor* with adaptive argumentation tutoring to write the review, the control group used the version of *ArgueTutor* without the adaptive argumentation tutoring, only receiving a general nonadaptive support based on the Toulmin theory (Toulmin 2003). We did not provide any introduction to any of the tools.

Posttest Phase: In the post-survey, we measured *perceived usefulness (***PU***), intention to use (***ITU***)* and *ease of use* (**PEOU**) following the technology acceptance model of (Venkatesh et al. 2003; Venkatesh and Bala 2008). Example items for the three constructs are: *"Imagine the tool was available in your next course, would you use it*?", *"The use of the tool enables me to write better argumentative texts."* or *"I would find the tool to be flexible to interact with"*. Moreover, we tested for the *perceived feedback accuracy (***PFA***)* (Podsakoff and Farh 1989) of both versions of *ArgueTutor* to control for the manipulation of the adaptive feedback by asking three items: "*The feedback I received reflected my true performance.", "The tool accurately evaluated my performance."* and *"The feedback I received from the tool was an accurate evaluation of my performance"*. Also, we asked the participants to judge their *perceived argumentation skill learning* (**PASL**), following the model of Toulmin (Toulmin 2003), by asking two items: "*I assume that the tool would help me improve my ability to argue in a structured way in texts." and "I assume that the tool would help me improve my ability to write more convincing texts.*" The answers were captured on a 1-to 5-point Likert scale (1: totally agree to 5: totally disagree, with 3 being a neutral statement). Additionally, 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?"*. Finally, we captured the demographics. In total, we asked 20 questions in the posttest.

Results

For data analysis, we performed a *double-sided t test* (*Welch's 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 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 group.

Group	n	ITU		PFA		PASL		PU		PEOU	
		Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
TG: adaptive support	22	2.4259	0.5088	2.6666	0.5362	2.1388	0.7030	2.4722	0.5808	1.7407	0.6720
CG: nonadaptive support	23	2.9365	0.8667	3.3015	0.6984	2.6190	0.6690	2.9285	0.9122	2.0317	0.7951
Total	45	*p = 0.02921 ** t-value = 2.2797		*p = 0.002788 *** t-value = -3.2066		*p = 0.03645 ** t-value = -2.1741		*p = 0.06738* t-value = -1.8889		*p = 0.2243* t-value = -1.238	

**p < 0.05, **p < 0.01*

Table 1. Measured mean and standard derivation on a 1-5 Likert scale (1: high, 5: low)

To answer our second research question, we compare the measured results between the five constructs we captured in the posttest between the two groups. Participants of the TG rated the ITU, the PFA and the PASL significantly higher than participants of the CG. A double-sided *t* test confirmed that the participants of TG, who received adaptive argumentation tutoring by *ArgueTutor*, perceived the ITU, the PFA and the PSAL of the PCA statistically significantly better than participants who only received general tutoring support (see Table 1: *ITU p < 0.05; PFA p < 0.01, PASL p < 0.05*). This indicates that from a student's perspective the value of *ArgueTutor* not only lies in the dialog-based interface and a step-by-step guidance but rather in the adaptivity of the PCA. Moreover, the results indicate that the overall perception of *ArgueTutor* is very positive. In both groups, the ITU, the PASL, the PU and the PEOU are higher than the neutral value of 3. Especially the perceived usefulness for writing better argumentative texts and the intention to use *ArgueTutor* 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 with. This will foster motivation and engagement to use the learning application. As described above, we also included open questions in our survey to receive the participants' opinions about the perception of the interaction of *ArgueTutor* to further evaluate our DF and DP. The general attitude of the interaction with *ArgueTutor* was very positive. Participants of the TG positively mentioned the fast and adaptive feedback (DP3 and DP4), the highlighting of argumentative components in their texts (DF5) and the adaptive feedback message on what to improve (DF6) several times. In contrast, almost every participant of the CG mentioned that individual feedback on the written text with a personal message on what to improve is a necessary feature, indicating again the importance of adaptivity of a PCA.

Step 8: Documenting the Evaluated Design Knowledge

To communicate our insights to the scientific knowledge base and to capture the results of our DSR project, we document our design knowledge according to the six core components of a design theory by Jones and Gregor (2007) (illustrated in Table 2). In this way, we summarize our theoretical contributions in the form of a *"design and action*" theory (Gregor and Jones 2007) based on our rigorously conducted design process. The purpose of the theory is to provide principles of form and function for constructing artifacts.

1) Purpose and scope	The purpose of *ArgueTutor* is to support students to learn how to argue by providing individual support, guidance and tutoring when solving argumentative writing tasks.
2) Constructs	**DF1-DF2:** Pedagogical Conversational Agent, **DF3-DF6:** writing support and individual writing feedback in learning mode, **DF7:** casual chat function, **DF5-DF6**: intelligent feedback algorithm based on NLP an ML.
3) Principles of form and function	**DP1**: PCA with guidance and support for argumentative writing. **DP2**: web-based PCA with responsive, simple and functional UX. **DP3**: proactive argumentative tutoring and guidance with explanations based on argumentation theory. **DP4**: analysis of individual argumentative components and individual feedback message. **DP5**: casual chat function, e.g., with jokes or fun facts.
4) Artifact mutability	Core design features, e.g. the AM algorithm, might be adapted to different pedagogical scenarios, e.g., content and language of texts. Human-like design elements of the PCA need to be adapted based on culture or domain

5) Testable propositions	(1) Using *ArgueTutor* increases the student's ability to write persuasive texts independent of a human instructor. (2) Using *ArgueTutor* improves the provision of guidance to students. (3) Using *ArgueTutor* reduces the required amount of tutoring to write argumentative texts from human instructor.
6) Justificatory knowledge	ICAP Framework (Chi and Wylie 2014), Toulmin Argumentation Model (Toulmin 2003)

Table 2. Documentation of our design knowledge adapted from Gregor and Jones (2007)

Discussion and Conclusion

In this DSR project, we followed Hevner (2007) to design, build and evaluate *ArgueTutor*, an adaptive dialog-based argumentation tutoring system that assists students on an individualized level in writing argumentative texts and supports their argumentation skill development independent of an instructor, time and place. We rigorously derived requirements from 85 scientific papers (rigor) and from twelve semi-structured user interviews (relevance) to formulate a concise set of design principles for adaptive PCAs for argumentation skill learning. We refined and evaluated those design principles as an instantiated artifact (*ArgueTutor)* in two design cycles with 77 users in total. We evaluated how students perceive the usefulness of our design principles for adaptive argumentation writing tutoring through *ArgueTutor* in an experiment with 45 participants. During this design process, we evaluated the results of each design iteration and successfully completed our design process after the two iterations.

Besides the software artifact as a situated implementation of an adaptive dialog-based learning system, we contribute design knowledge to the scientific knowledge base. We systematically deduced design knowledge as documented in our last step of the design research (Table 2 in step 8). Due to the systematic procedure, we aimed at generating a satisfying design contribution (Gregory and Muntermann 2014). The resulting design knowledge is not only valid for our specific case but can also be transferred to further use cases in adaptive argumentation learning. For instance, it is easily possible to apply the concept of *ArgueTutor* in courses that deal with other content or other languages. For this purpose, only the back end algorithm needs to be adapted to the other scenario. As described, multiple corpora and AM models exist in literature that can be easily embedded in *ArgueTutor*, e.g., for English student essays (Stab and Gurevych 2017) or English law cases (Mochales Palau and Ieven 2009). The design principles of form and function and the overall system design do not need to be adapted for those use cases. Furthermore, it is also possible to transfer the design knowledge to pedagogical scenarios that target the training of other metacognition skills. For instance, if the learning of general feedback skills or empathy skills of students is aimed to be trained, a similar PCA could be used. However, in this case, the system design might need to be revised partially, e.g., the argumentation feedback in the text and the theory model need to be adapted. Due to this transferability of our design knowledge, our research does not only provide a Level 1 DSR contribution by showing a situated artifact implementation but also provides a nascent design theory (Level 2 contribution) (Gregor and Hevner 2013). In addition to these contributions to the scientific knowledge base, our results are also valuable for the practical use of dialog-based technologies, such as NLP, ML or AM, for PCAs in education. Our results indicate that the state of the art in NLP and ML techniques are suited to design complex skill tutoring systems that are able to support students individually – even in existing pedagogical scenarios such as writing persuasive texts. We have thus shown that more advanced support possibilities in skill education can be implemented using PCAs in contrast to common technology-enhanced approaches.

However, our research also faces several 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. We focused on deducing and evaluating design principles and assessing the students' evaluation. Besides, novelty effects of students using our PCA for the first time cannot be expelled. For future research, particularly analyzing the long-term effect of using *ArgueTutor*, we aim to implement the artifact into our learning management system and measure long-term effects on usability and acceptance of such a 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. 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

Agarwal, R., and Karahanna, E. 2000. "Time Flies When You're Having Fun: Cognitive Absorption and Beliefs about Information Technology Usage," *MIS Quarterly* (24:4), p. 665.

Ashford, S. J. 1986. "Feedback-Seeking in Individual Adaptation : A Resource Perspective," *Academy of Management Journal* (29:3), pp. 465–487.

Atkinson, R. C., and Shiffrin, R. M. 1968. "Human Memory: A Proposed System and Its Control Processes," *Psychology of Learning and Motivation - Advances in Research and Theory* (2:C), pp. 89–195.

von Aufschnaiter, C., Erduran, S., Osborne, J., and Simon, S. 2008. "Arguing to Learn and Learning to Argue: Case Studies of How Students' Argumentation Relates to Their Scientific Knowledge," *Journal of Research in Science Teaching* (45:1), pp. 101–131.

Bird, S., Klein, E., and Loper, E. 2009. "Natural Language Processing with Python," *Text* (Vol. 43).

Black, P., and Wiliam, D. 2009. "Developing the Theory of Formative Assessment," *Educational Assessment, Evaluation and Accountability* (21:1), pp. 5–31. (

Bocklisch, T., Faulkner, J., Pawlowski, N., and Nichol, A. 2017. *Rasa: Open Source Language Understanding and Dialogue Management.*

Boltuži, F., and Šnajder, J. 2014. "Back Up Your Stance : Recognizing Arguments in Online Discussions," *Proceedings of the First Workshop on Argumentation Mining*, pp. 1–43.

vom Brocke, J., Maaß, W., Buxmann, P., Maedche, A., Leimeister, J. M., and Pecht, G. 2018. "Future Work and Enterprise Systems," *Business and Information Systems Engineering* (60:4), pp. 357–366.

vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., and Cleven, A. 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).

Cai, W., Grossman, J., Lin, Z., Sheng, H., Tian, J., Wei, Z., Williams, J. J., and Goel, S. 2019. *MathBot: A Personalized Conversational Agent for Learning Math.*

Cameron, G., Cameron, D., Megaw, G., Bond, R., Mulvenna, M., O'Neill, S., Armour, C., and McTear, M. 2017. "Towards a Chatbot for Digital Counselling," *HCI 2017: Digital Make Believe - Proceedings of the 31st International BCS Human Computer Interaction Conference.*

Chi, M. T. H., and Wylie, R. 2014. "The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes," *Educational Psychologist* (49:4), pp. 219–243.

Cohn, M. 2004. "User Stories Applied For Agile Software Development."

Cooper, H. M. 1988. "Organizing Knowledge Syntheses: A Taxonomy of Literature Reviews," *Knowledge in Society* (1:1), pp. 104–126.

Driver, R., Newton, P., and Osborne, J. 2000. "Establishing the Norms of Scientific Argumentation in Classrooms," *Science Education* (84:3), John Wiley & Sons, Ltd, pp. 287–312.

Duschl, R. A., and Osborne, J. 2002. "Supporting and Promoting Argumentation Discourse in Science Education," *Studies in Science Education* (38:1), pp. 39–72.

Dusmanu, M., Cabrio, E., and Villata, S. 2018. "Argument Mining on Twitter: Arguments, Facts and Sources," *Proceedings of the 2017 Conference on Empirical Methods in NLP* pp. 2317–2322.

Eemeren, F. H. van, Grootendorst, R., Johnson, R. H., Plantin, C., Willard, C. A., Grootendorst, R., Johnson, R. H., Plantin, C., and Willard, C. A. 1996. *Fundamentals of Argumentation Theory*, Routledge.

Eger, S., Daxenberger, J., and Gurevych, I. 2017. "Neural End-to-End Learning for Computational Argumentation Mining," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 11–22.

Fadel, C., Bialik, M., and Trilling, B. 2015. *Four-Dimensional Education : The Competencies Learners Need to Succeed.*

Flender, J., Christmann, U., and Groeben, N. 1999. "Entwicklung Und Erste Validierung Einer Skala Zur Erfassung Der Passiven Argumentativ-Rhetorischen Kompetenz," *Zeitschrift Für Differentielle Und Diagnostische Psychologie* (20:4), Verlag Hans Huber, pp. 309–325.

Fromm, H., Wambsganss, T., and Söllner, M. 2019. "Towards a Taxonomy of Text Mining Features," in *European Conference of Information Systems (ECIS)*, pp. 1–12.

Gläser, J., and Laudel, G. 2010. *Experteninterviews Und Qualitative Inhaltsanalyse : Als Instrumente Rekonstruierender Untersuchungen*, VS Verlag für Sozialwiss.

Gregor, S., Chandra Kruse, L., and Seidel, S. 2020. “The Anatomy of a Design Principle,” *Journal of the Association for Information Systems* (Forthcomin).

Gregor, S., and Hevner, A. R. 2013. *Positioning and Presenting Design Science Research for Maximum Impact*.

Gregory, R. W., and Muntermann, J. 2014. “Research Note: Heuristic Theorizing: Proactively Generating Design Theories,” *Information Systems Research*, INFORMS, pp. 639–653.

De Groot, R., Drachman, R., Hever, R., Schwartz, B., Hoppe, U., Harrer, A., De Laat, M., Wegerif, R., Mclaren, B. M., and Baurens, B. 2007. “Computer Supported Moderation of E-Discussions: The ARGUNAUT Approach.”

Hattie, J., and Timperley, H. 2007. “The Power of Feedback,” *Review of Educ. Research* (77:1), pp. 81–112.

Hevner, A. R. 2007. “A Three Cycle View of Design Science Research,” *Scandinavian Journal of Information Systems*.

Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. “Design Science in Information Systems Research,” *Design Science in IS Research MIS Quarterly* (28:1), p. 75.

Hobert, S. 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, Munich, Germany*.

Hobert, S., and Wolff, R. M. Von. 2019. “Say Hello to Your New Automated Tutor – A Structured Literature Review on Pedagogical Conversational Agents,” *14th International Conference on Wirtschaftsinformatik, Siegen, Germany*.

Hu, T., Xu, A., Liu, Z., You, Q., Guo, Y., Sinha, V., Luo, J., and Akkiraju, R. 2018. “Touch Your Heart: A Tone-Aware Chatbot for Customer Care on Social Media,” in ACM CHI *Conference on Human Factors in Computing Systems*.

Jonassen, D. H., and Kim, B. 2010. “Arguing to Learn and Learning to Argue: Design Justifications and Guidelines,” *Educational Technology Research and Development* (58:4), pp. 439–457.

Jones, D., and Gregor, S. 2007. “The Anotomy of a Design Theory,” *Journal of the Association for Information Systems* (8:5), pp. 312–335.

Krassmann, A. L., Paz, F. J., Silveira, C., Tarouco, L. M. R., and Bercht, M. 2018. “Conversational Agents in Distance Education: Comparing Mood States with Students’ Perception,” *Creative Education* (09:11), pp. 1726–1742.

Kuhn, D. 1992. “Thinking as Argument,” *Harvard Educational Review* (62:2), pp. 155–179.

Kuhn, D. 1993. “Science as Argument: Implications for Teaching and Learning Scientific Thinking,” *Science Education* (77:3), John Wiley & Sons, Ltd, pp. 319–337.

Kuhn, D. 2005. *Education for Thinking*, Harvard University Press.

Kulik, J. A., and Fletcher, J. D. 2016. “Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review,” *Review of Educational Research* (86:1), pp. 42–78.

Lawrence, J., and Reed, C. 2019. “Argument Mining: A Survey,” *Compu. Linguistics* (45:4), pp. 765–818.

Lippi, M., and Torroni, P. 2016. “MARGOT: A Web Server for Argumentation Mining,” *Expert Systems with Applications* (65), Elsevier Ltd, pp. 292–303

Mochales Palau, R., and Ieven, A. 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,* pp. 21–30.

Moens, M., Boiy, E., and Reed, C. 2007. “Automatic Detection of Arguments in Legal Texts,” *Proceedings of the 11th International Conference on Artificial Intelligence and Law*.

Nass, C., and Moon, Y. 2000. “Machines and Mindlessness: Social Responses to Computers,” *Journal of Social Issues* (56:1), John Wiley & Sons, Ltd (10.1111), pp. 81–103.

Newton, R. P., Roef, L., Witters, E., and Van Onckelen, H. 1999. “Tansley Review No. 106,” *New Phytologist* (143:3), John Wiley & Sons, Ltd (10.1111), pp. 427–455.

OECD. 2018. *The Future of Education and Skills - Education 2030*.

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., and Yao, S. Y. 2016. “The Development and Validation of a Learning Progression for Argumentation in Science,” *Journal of Research in Science Teaching* (53:6), pp. 821–846.

Payr, S. 2003. “The Virtual University’s Faculty: An Overview of Educational Agents,” *Applied Artificial Intelligence* (17:1), Taylor & Francis Group , pp. 1–19.

Pinkwart, N., Ashley, K., Lynch, C., and Aleven, V. 2009. “Evaluating an Intelligent Tutoring System for Making Legal Arguments with Hypotheticals,” *International Journal of AI in Eduducation* (Vol. 19).

Podsakoff, P. M., and Farh, J. L. 1989. “Effects of Feedback Sign and Credibility on Goal Setting and Task

Performance," *Organizational Behavior and Human Decision Processes* (44:1), pp. 45–67.

Ruan, S., Jiang, L., Xu, J., Tham, B. J.-K., Qiu, Z., Zhu, Y., Murnane, E. L., Brunskill, E., and Landay, J. A. 2019. *QuizBot: A Dialogue-Based Adaptive Learning System System for Factual Knowledge*, CHI.

Scheuer, O. 2015. *Towards Adaptive Argumentation Learning Systems*.

Scheuer, O., Mclaren, B. M., Loll, F., and Pinkwart, N. 2012. "Automated Analysis and Feedback Techniques to Support and Teach Argumentation: A Survey," *Educational Technologies for Teaching Argumentation Skills* (2), pp. 71–124.

Seaman, J. E., Allen, I. E., and Seaman, J. 2018. "Higher Education Reports".

Shawar, B. A., and Atwell, E. S. 2005. "Using Corpora in Machine-Learning Chatbot Systems," *International Journal of Corpus Linguistics* (10:4), pp. 489–516.

Soloway, E., Jackson, S. L., Klein, J., Quintana, C., Reed, J., Spitulnik, J., Stratford, S. J., Studer, S., Eng, J., and Scala, N. 1996. "Learning Theory in Practice: Case Studies of Learner-Centered Design."

Song, Y., Heilman, M., Klebanov, B. B., and Deane, P. 2014. "Applying Argumentation Schemes for Essay Scoring."

Stab, C., and Gurevych, I. 2014. "Identifying Argumentative Discourse Structures in Persuasive Essays," in *Conference on Empirical Methods in Natural Language Processing*, pp. 46–56.

Stab, C., and Gurevych, I. 2017. "Parsing Argumentation Structures in Persuasive Essays," *Computational Linguistics* (43:3), pp. 619–659.

Stenetorp, P., Pyysalo, S., and Topi, G. 2012. *BRAT : A Web-Based Tool for NLP-Assisted Text Annotation*.

Suppes, P., and Morningstar, M. 1969. "Computer-Assisted Instruction," *Science* (166:3903), pp. 343–350.

Suthers, D. D., and Hundhausen, C. D. 2001. "European Perspectives on Computer-Supported Collaborative Learning."

Topi, H. 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*.

Toulmin, S. E. 2003. "The Uses of Argument: Updated Edition," *The Uses of Argument: Updated Edition*.

Venable, J., Pries-Heje, J., and Baskerville, R. 2016. "FEDS: A Framework for Evaluation in Design Science Research," *European Journal of Information Systems* (25:1), Nature Publishing Group, pp. 77–89.

Venkatesh, V., and Bala, H. 2008. "Technology Acceptance Model 3 and a Research Agenda on Interventions," *Decision Sciences* (39:2), John Wiley & Sons, Ltd (10.1111), pp. 273–315.

Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. 2003. "User Acceptance of Information Technology: Toward a Unified View," *MIS Quarterly* (27:3), pp. 425–478.

Vygotsky, L. S. 1980. *Mind in Society: The Development of Higher Psychological Processes*, Harvard UP.

Walton, D., Reed, C., and Macagno, F. 2008. *Argumentation Schemes*, Cambridge: Cambridge University

Wambsganss, T., Molyndris, N., and Söllner, M. 2020. "Unlocking Transfer Learning in Argumentation Mining: A Domain-Independent Modelling Approach," in *15th International Conference on Wirtschaftsinformatik*, Potsdam, Germany.

Wambsganss, T., Niklaus, C., Cetto, M., Söllner, M., Leimeister, J. M., and Handschuh, S. 2020. "AL : An Adaptive Learning Support System for Argumentation Skills," in P*roceedings of the 2020 CHI Conference on Human Factors in Computing Systems, New York, NY, USA: ACM*.

Wambsganss, T., and Rietsche, R. 2020. "Towards Designing an Adaptive Argumentation Learning Tool," in *40th International Conference on Information Systems,* Munich, Germany.

Wambsganss, T., Winkler, R., Schmid, P., and Söllner, M. 2020a. "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*.

Wambsganss, T., Winkler, R., Schmid, P., and Söllner, M. 2020b. "Designing a Conversational Agent as a Formative Course Evaluation Tool," in *15th International Conference on Wirtschaftsinformatik*, Potsdam, Germany.

Wambsganss, T., Winkler, R., Söllner, M., and Leimeister, J. M. 2020. "A Conversational Agent to Improve Response Quality in Course Evaluations," in P*roceedings of the 2020 CHI Conference on Human Factors in Computing Systems, New York, NY, USA: ACM.*

Winkler, R., Büchi, C., and Söllner, M. 2019. "Improving Problem-Solving Skills with Smart Personal Assistants: Insights from a Quasi Field Experiment," in *40th International Conference on Information Systems, Munich, Germany.*

Winkler, R., and Söllner, M. 2018. "Unleashing the Potential of Chatbots in Education : A State-Of-The-Art Analysis." In Academy of Management, *A O M*, Chicago, USA.