Universidad del País Vasco Euskal Herriko Unibertsitatea

First Steps towards an Argumentative Dialogue System for Oracy Skills Development in Schools

Author: Cecilia Domingo

Advisors: Nora Aranberri & Arantxa Otegi

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua
Language Analysis and Processing

Final Thesis

June 2021

Departments: Computer Systems and Languages, Computational Architectures and Technologies, Computational Science and Artificial Intelligence, Basque Language and Communication, Communications Engineer.

Abstract

There is academic and institutional interest in implementing dialogic approaches to teaching, as dialogue has been shown to be both a means towards learning and a goal in a global society where oracy skills are crucial. Some technologies have already proven useful in dialogic teaching. However, no link has been established between dialogic teaching and dialogue systems, which might be suitable for this approach. We thus analyze the literature on dialogic teaching, dialogue systems for education, and dialogue-system design to describe the features that could result in a successful argumentative dialogue system for dialogic teaching. We propose an example system and compile a dataset of possible student input. Tests performed on the dataset reveal that basic text annotation and a semantic similarity tool may be useful for assessing students' performance on the analytical part of the proposed task; further data gathering and tests are needed to draw conclusions for the remainder of the task.

Keywords:

dialogue system, chatbot, dialogic teaching, oracy skills

Contents

List of Figures

List of Tables

1 Introduction

Oracy skills[1] have been shown to impact people's career prospects and social status. Moreover, they are also important during their learning process. These skills enable students to participate in classroom dialogue, which has been linked to cognitive development and increased learning gains. Therefore, the development of these skills is both a goal and a means for overall learning in dialogic teaching[2] (Alexander, 2010; Mercer et al., 2010a; Mercer and Howe, 2012; Mercer et al., 2017, 2019).

Despite the interest in implementing this teaching approach (Jay et al., 2017; Sedova, 2017; Mercer et al., 2019), there are several obstacles that make it difficult: teachers are given little time to cover content-dense curricula, the number of students is often high, students' skills may differ significantly, some practices and ideas that clash with the dialogic approach may be too deeply ingrained in the learning culture... (Mercer et al., 2010a, 2017; Sedova, 2017; Okada et al., 2018). Technology can help overcome some of these obstacles (Mercer et al., 2010a; Mercer and Howe, 2012; Major et al., 2018).

Studies linking the use of technology with dialogic teaching focus on tools like interactive whiteboards, micro-blogging sites or online forums (ibid), but none look at the potential of one tool which might seem ideally suited for a teaching approach based on dialogue: dialogue systems. These tools have already shown potential for enhancing student learning, in ways such as aiding memorization (Abbasi and Kazi, 2014) or improving student participation in discussions (Goda et al., 2014). Given this lack of dialogue systems specifically designed for dialogic teaching, in this project we have attempted to fill this gap by defining which characteristics would be necessary in a dialogue system to implement dialogic teaching. Often what impedes the development of dialogue systems that can be used in education is a lack of data and technical resources (Kuyven et al., 2018). Nonetheless, with this project we aim to at least provide guidelines as well as a concrete example to ease the process of designing such systems and to ensure that the resulting tool has the pedagogical motivation needed for its success in an educational context.

We begin our work with the hypothesis that, given that both dialogic teaching and dialogue systems have dialogue as their core, dialogue systems can be helpful as teaching and learning tools in this pedagogic approach. Studies linking other technologies with dialogic teaching, like the work of Major et al. (2018), support this intuition which will be tested through a literature review. As, to our knowledge, there is currently no dialogue system designed specifically with dialogic teaching in mind, we will not test the usefulness of such tools directly on a student sample. Instead, we will endeavor to extract from the literature the main requisites of dialogic teaching, analyze which features have already been demonstrated to aid dialogic teaching with other technologies, and evaluate whether these requisites and features could be included in the design of a dialogue system. The goal of this

[1]communication skills, in the broadest sense of the word, which includes physical subskills, linguistic subskills, cognitive subskills and social and emotional subskills (Mercer et al., 2017). A more detailed definition can be found in sections 1.2 and 2.2.1.

[2]Teaching approach that considers dialogue as a tool for knowledge construction. This definition is elaborated on sections 1.2 and 2.1.1.

project is to indicate which dialogue system features might be necessary for the system to be used in dialogic teaching. This description of dialogue system features linked to dialogic teaching is presented as a theoretical framework (section 4). However, this project also includes additional contributions beyond the primary goal. Two important sets of subskills needed in dialogic teaching and reinforced by it are cognitive skills and social skills. A detailed description of these skills is given in sections 1.2 and 2.2.1. The nature of these skills makes reasoned discussion a key tool in dialogic teaching (Alexander, 2010; Mercer et al., 2017). For this reason, our literature review also explores dialogue systems designed for debating, even if they were not created for educational purposes (section 3.3). We also provide an example of a concrete realization of our proposed framework for the design of a dialogue system that can be used for dialogic teaching - our proposed task takes the form of an argumentation task (section 5). In addition to a description of our proposed task, we include an example imagined conversation (Appendix B). Additionally, we perform some initial tests on its feasibility and technical requirements (section 6.2). To perform these tests, a small dataset is compiled with the type of data needed for one part of the task (section 6.1)[3]. Our tests, as we detail through section 6.2, suggest that simple tools and data like the dataset we compiled might suffice for an initial sketch of a simplified dialogue system that can be progressively improved to arrive at a system that is useful and effective for dialogic teaching.

1.1 Document structure

We will finish this introductory section with a subsection defining the key concepts that our project deals with (section 1.2). We then carry out a review of the literature concerning the pedagogical approach that we explore in this project (section 2). We first define the approach, summarizing the main theoretical aspects (section 2.1). After that, we explain why the relevance of this pedagogical approach and why, despite its promise, it is not more widely implemented (section 2.2). We finish the pedagogical review by explaining how technology may facilitate the implementation of dialogic teaching (section 2.3).

The pedagogical review is followed by a review of the literature on dialogue systems (section 3). We provide a general description of what these systems are (section 3.1) and how they are designed (section 3.2). We then comment on argumentative dialogue systems, a specific type of dialogue system which we believe most linked to dialogic teaching (section 3.3).

Our first main contribution can be found in section 4. There, we provide the framework linking dialogic teaching with features of a dialogue system ideally suited for this pedagogic approach. Afterwards, we propose a specific task that follows our framework, in the form of an argumentative task (section 5). That section, in addition to discussing several design aspects, introduces an example of how the task might be carried out (section 5.7 and Appendix B). It also includes a detailed analysis of how the proposed task aligns with our framework (section 5.8). Then, in section 6, we describe how we perform initial tests

[3]The dataset can be accessed at this link: SAT Dataset

on what the proposed system might require (i.e. data and tools). For this, we compile a dataset, described in section 6.1, and use it to carry out an experiment, described in section 6.2. We finish with a section summarizing our conclusions and acknowledging the limitations we encountered (section 7)

We also include three appendices. The first one is a list of the dialogue acts of the proposed task. What these acts entail is introduced in section 3.1 and further described in section 5.6. The second appendix is an example of the form that the task proposed in section 5 might take. Section 5.7 provides more details. Finally, the third appendix is an example of what is contained in the dataset[4]. This example is an annotated text, and this is the text used as reference for the example conversation from Appendix B. Throughout this document we sometimes use the text from Appendix C as an example. However, for certain aspects we have found better examples in other texts; still, we introduce all our examples with a description of the main ideas of the text that they come from, so that readers can follow our examples without needing to look at the whole dataset. We also believe that including a variety of examples illustrates the diversity of topics covered in the dataset.

1.2 Definitions

Before we can continue to discuss this project, it seems necessary to define three key concepts, especially considering the multidisciplinary nature of this work. This project deals with two key concepts from the field of education: dialogic teaching and oracy skills. The first concept, dialogic teaching, is a pedagogic approach, meaning that it is not linked to any specific teaching method, but rather describes a philosophy that can be applied with a broad variety of techniques (Alexander, 2010). This approach has talk at its core, drawing from the theories of Bathkin and Vygotsky, where language and talk are seen as tools to build knowledge and thought (Mercer and Howe, 2012; Mercer et al., 2019). The details of what this approach entails are further explained in section 2.1; the main ideas, however, can be summarized as teaching that makes students "think, not just report someone else's thinking" (Alexander, 2010, p. 4) and where teachers talk *with* the students, instead of *at* the students (Skidmore, 2019, in Mercer et al., 2019).

As talk is crucial in dialogic teaching, "oracy skills" are also an important concept, as these are the skills that enable students to better participate in this type of teaching as well as the skills that this approach reinforces (Alexander, 2010; Mercer et al., 2010a). Oracy skills can be understood as communication skills, but not in the simplistic sense that this word has come to acquire, of being able to listen and speak in a language; rather, they are a very broad set of subskills that people use to "express their thoughts and communicate with others" (Alexander, 2012, in Mercer, 2017, p. 52). These subskills are further explained in section 2.2.1.

The third key concept that we explain here, this time from a more technological field, is "dialogue systems", also known as "chatbots"; we will use the former term, as the latter

[4]The complete dataset can be accessed at this link: SAT Dataset

is not always synonymous with "dialogue system", but rather one of its types (Jurafsky and Martin, 2019). Dialogue systems can be defined as programs that "communicate with users in natural language" (Jurafsky and Martin, 2019). This broad definition includes two types of programmes: task-oriented systems and non-task oriented systems, more commonly known as chatbots (Chen et al., 2018; Jurafsky and Martin, 2019). As the name suggests, task-oriented systems are programs that employ natural language to help users carry out a task, such as obtaining a weather forecast; chatbots, on the other hand, converse merely to entertain or accompany users or to make interactions with task-oriented systems more natural (ibid).

2 Review of pedagogical literature

In this section we firstly review the literature on what constitutes dialogic teaching. On the second part of the review, we provide references to justify the relevance of this teaching approach; we then explain why this approach, despite the research supporting it, has not been more widely implemented. On the third part, we explain how technology might help in the implementation of dialogic teaching and then examine the ways in which dialogue systems specifically are being used in education.

2.1 What is dialogic teaching?

2.1.1 Dialogic teaching principles

As was briefly explained in the Definitions section (1.2), dialogic teaching is a pedagogic approach (i.e. not a method, but rather a philosophy that can be reflected in very different practices) where the focus is on talk. This is not to say that talk is not already present in most teaching practice, but dialogic teaching moves away from traditional monologic practices where teachers speak and the students' role is only to listen to answer the occasional closed question (Alexander, 2010); in dialogic teaching, students are active participants (ibid). The reasoning for this approach stems from Vygotsky and Bathkin's theories that language is our tool for knowledge construction both in our minds and through interaction with others: truth is found via dialoguing with others, and to integrate that truth into our understanding of the world, we must put it into our words (i.e. our own voice) (Vygotsky, 1978, in Mercer, 2010). These philosophical theories are translated into the five principles that define dialogic teaching; talk, in order to be truly dialogic and conducive to learning, needs to be collective, reciprocal, supportive, cumulative and purposeful (Alexander, 2010).

Collectivity: this principle means that there is more than one participant: it is not a monologue where only one person speaks; instead, two or more people may intervene (ibid).

Reciprocity: this is very closely related to the first principle; it is not enough that more than one person participate, they also have to be active speakers, as well as active listeners (Skidmore, in Mercer et al., 2019).

Supportiveness: all participants feel that they are allowed to contribute, that they will be listened to, and that all participants will cooperate to understand each other (Alexander, 2010).

Cumulativeness: ideas have to be built upon; they need to be analyzed and evaluated, and questions need to lead to further questions for further knowledge construction (ibid).

Purposefulness: dialogue, to be pedagogical, needs to have a purpose. Though students' ideas may sometimes lead to topics outside the curriculum which are worth exploring, classroom activities need to be designed to reach a desired learning goal (ibid).

2.1.2 Scaffolding

It is important to bear in mind that dialogic teaching is a teaching approach; therefore, while the dialogic aspect warrants attention, the teaching aspect is also important. This implies that there not only needs to be dialogue in general, but also scaffolded dialogue (Alexander, 2010). Scaffolding is a type of guidance where supportive dialogue is used to lead the student through the steps of a problem and, as the student progresses, the support becomes less necessary and is gradually removed (Mitchell et al., 2013). As much of the basis of dialogic teaching, this idea stems from Vygotsky's theories, especifically that learning involves moving from "other-regulation" to "self-regulation" and from "inter-mental" activity to "intra-mental" activity: we build knowledge by interacting with others, until we assimilate it into our own voice and it becomes our knowledge (Vygotsky, 1975, in Mitchell et al., 2013).

2.1.3 Dialogic teaching indicators

Alexander's principles of dialogic teaching, though influential, are deeply rooted in theory, so they may not be useful for determining whether dialogic teaching is taking place in a class (Sedova, 2017). Principles have thus been linked to indicators that are easier to analyze (ibid). Different authors have analyzed the implementation of dialogic teaching in different ways, sometimes focusing on the learning outcomes (Jay et al., 2017), assuming that dialogic teaching is really taking place. Sedova (2017) proposes a set of five indicators that were effectively used to assess the implementation of dialogic teaching in a longitudinal study, where the indicators could be used to analyze individual lessons and observe the progression of the implementation process (i.e. seeing how dialogic each class was and how more or less dialogic they became). The five indicators are: students reason their ideas, teachers ask open questions that require students to think, there is uptake (i.e. ideas are followed by new questions or contributions that expand upon that idea), and there is open discussion (i.e. there are several participants each talking for at least thirty seconds) (ibid).

Parallel to Alexander's principles and the ensuing indicators, Mercer et al. (2010b) also offer some signs of dialogic teaching taking place which are based on observations of students participating in collaborative tasks (i.e. where a group of students work towards a shared goal). One first important conclusion that can be extracted from these observations (ibid) is that putting students in a group does not mean that there will be collaborative work - for there to be successful dialogue, some rules need to be set and enforced (Alexander, 2010). This is linked to the main conclusion of the observations, where Mercer et al. (2010b) describe the three broad types of talk that the students engaged in and what talk type is conducive to learning (and should be fostered) and which one hinders learning (and should be discouraged). The three types of talk are: exploratory talk, cumulative talk (not to be confused with Alexander's principle), and disputational talk (ibid).

Exploratory talk is what was linked to dialogic teaching, as it involves students expressing ideas in a reasoned manner and building up on each other's views - all in line with Alexander's principles. Mercer et al. (2010b) also observed what they called cumulative talk, where students simply revisit what has been discussed so far, seeing which ideas lead to which other ideas and which ones were agreed upon. This in itself is not conducive to learning, but it is a type of talk that Mercer et al. (2010b) observed frequently and considered positive, as it is linked to the important exploratory talk, and it is especially useful when students need to summarize their discussion to link their conclusions to the task goal. The remaining type of talk that Mercer et al. (2010b) observed is disputational talk, which is the type of talk that should be avoided (e.g. by enforcing rules that discourage it, such as "All ideas need to be reasoned"). In disputational talk, students are not respectful of each other's contributions (there is no supportiveness), they are monologic (they monopolize the task and/or not listen to other participants), and/or they do not reason their ideas (there is no cumulativeness, as conclusions do not stem from other ideas).

2.2 Importance of dialogic teaching

Dialogic teaching has been the subject of extensive research (Major et al., 2018), and efforts are being made to implement this teaching approach in schools (Sedova, 2017). This is due to two broad factors, further explained in the following subsections: firstly, learning to use dialogue, the basis of dialogic teaching, is an important life skill (Mercer et al., 2017); secondly, dialogue is a tool that can help learn other skills more effectively (Jay et al., 2017; Major et al., 2018).

2.2.1 Dialogue as an end

As was explained in the Definitions section (1.2), the skills that dialogic teaching employs and reinforces are called oracy skills (Mercer et al., 2017). In order to better understand why the acquisition of these skills is considered important, we need to first provide a more detailed definition of what these are. As we mentioned, oracy skills are communication skills understood as a large set of subskills, rather than simply speaking and listening (ibid). These subskills can be classified into four groups: physical skills, cognitive skills, linguistic skills, and social and emotional skills (ibid). Physical skills cover the use of voice and body language to communicate; cognitive skills cover the choice of content, clarifying and summarizing ideas, reasoning and analyzing views, managing time while we communicate, maintaining focus, and being aware of our interlocutors' understanding; linguistic skills cover the use of vocabulary, grammar, register, discourse structure and rhetorical techniques to communicate; social and emotional skills, finally, cover managing interaction and turns, listening actively, and speaking confidently.

Possessing this wide range of skills has been linked to better employability and a higher social status (ibid). Improving students' oracy skills also helps them participate more actively in class (ibid). For these reasons, it seems advisable to make improving students' oracy skills an important goal in their education. This is especially the case for cultures that have traditionally given less importance to these skills, as neglecting students' oracy skills development puts them at a disadvantage in a global society (Okada et al., 2018).

2.2.2 Dialogue as means

The development of oracy skills is not the only goal of dialogic teaching; this approach can be used to teach different skills and curriculum topics (Jay et al., 2017; Major et al., 2018). Aside from helping achieve the specific goals of a course or lesson, dialogic teaching also has broader cognitive benefits (Major et al., 2018). For example, higher oracy skills have been linked to higher problem-solving skills (Azmitia & Montgomery, 1993, in Mercer & Howe, 2012; Underwood & Underwood, 1999, in Mercer & Howe, 2012). Dialogic teaching has also been linked to improved reading skills: in interventions aimed at improving literacy, the best outcomes were achieved when students were encouraged to express ideas in their own words, instead of completing prompts or answering yes/no questions (Wolf et al., 2006, in Mercer & Howe, 2012).

2.2.3 Challenges for implementation

Research mentioned in the previous sections strongly suggests that dialogic teaching is a beneficial approach (Jay et al., 2017). However, its implementation is not easy (Sedova, 2017). One of the challenges is simply a lack of time, as packed curricula may leave little or no time for student participation (Mercer et al., 2010a). Another difficulty is that students' oracy skills may be very heterogeneous, making it difficult for all of them to participate in dialogic activities in class (Mercer et al., 2017). Students may be exposed to the same content at school, but outside the classroom their environments may be very different, and some may have no model of dialogue to imitate in their daily life (Mercer et al., 2017).

As significant as these obstacles are, one more challenging factor seems to be tradition and attitude (Mercer et al., 2010a; Sedova, 2017). Implementing a dialogic approach requires teachers (as well as students and administrators) reshaping their views and practices to understand the benefits of this approach and how it is to be applied (Mercer et al., 2010a). Making classes more dialogic may involve significant changes for teachers used to different approaches; if teachers lack guidance to adapt their methodology to this approach, they may sometimes resort to old methods that are not in line with dialogic principles (Sedova, 2017). Even when teachers are offered formative programs to understand the principles of dialogic teaching, this is not enough if the programs do not include adequate feedback that allows teachers to reflect on their practice (Sedova, 2017). This necessary training for teachers, as well as additional planning and reviewing, implies an investment of time and money (Jay et al., 2017).

2.3 Technology for dialogic teaching

Technology can help teachers and students overcome some challenges to the implementation of dialogic teaching, as demonstrated most significantly by the exhaustive literature review carried out by Major et al. (2018), which covers 71 studies on this topic. The first benefit of technology for dialogic teaching that is observed in this review is that technology helps students access each other's ideas and build cumulative knowledge (e.g. an internet forum can store contributions from a large number of students, exposing them to a wider variety of views than could be covered in a traditional in-class discussion). The traceability of ideas also improves continuity across lessons (e.g. a computer can give access to documents from previous lessons and show who contributed what and when) (ibid). At the same time that previous ideas are visible, students can change and expand them: for example, while, for traceability, users may access the log of changes made to a document, they are also able to easily edit digital content. This gives students flexibility and freedom and allows them to test ideas without fear of not having the correct answer (Mercer et al., 2010b; Major et al., 2018).

Technological tools also allow learning to occur outside the classroom, which is especially useful for students lagging behind or in situations where in-person learning is not possible (Major et al., 2018). Students in distance education report feeling isolated; also students in flipped classrooms, where only part of the work is done alone at home, miss having feedback during their autonomous work (Huang et al., 2019). Technology can tackle these issues by providing students with instant feedback (e.g. quizzes that are automatically corrected), or adding new and sometimes more welcoming channels of communication with peers and teachers (e.g. a forum where questions can be asked anonymously or a chat where students can help each other). Additionally, technology can help teachers monitor students' work more closely, even in distance education (e.g. a log of students' work can be available to teachers, showing them where students had to try more times to get the right answer on a quiz, or which tasks took longer to complete); in class, this information about students' progress can also be made accessible to teachers, enabling them to provide better formative feedback without having to loom over the students, which might hinder their work or, in large classrooms, leave some students unattended (Mercer et al., 2010b; Major et al., 2018).

Whether in a classroom or in distance learning, technology that facilitates dialogic teaching can also improve the relationship between teacher and students and among students (Major et al., 2018). One specific way in which this can happen is that participation may increase, possibly as a result of the motivating effect of interesting technology, due to the supportive environment created by technology, or by the ability of technology to simulate discussions to prime students for that type of task (Goda et al., 2014; Major et al., 2018). Goda et al. (2014) observed this specifically with chatbots: students who used a chatbot for practice before a discussion participated more actively in it.

Another way that technology aids learning, as observed by Major et al. (2018) is its potential multimodality , which allows teachers to design lessons that appeal to different learning styles, as well as improve accessibility. Multimodality can mean simply combining written and spoken communication, but also gestural communication in the case of systems with an avatar or robots (Heller, 2016; Thies et al., 2017; Marge et al., 2020). Multimodality is specifically recommended for dialogue systems, as it makes communication more natural and increases the user's freedom (Jokinen, 2009). Jokinen (2009) also advocates for chatbot design that has accessibility in mind, as this then proves to benefit not only users with special needs.

Lastly, it is important to emphasize that any technology that is to assist in the implementation of dialogic teaching needs to be specifically designed with this approach in mind (Mercer et al., 2010b; Major et al., 2018). Many technological tools used in education are designed and implemented without a clear pedagogical purpose in mind, which leads to their potential being unfulfilled or students' and teacher's needs not being met by the new tools (ibid).

2.3.1 Dialogue systems used in education

Some tools that were not specifically designed for dialogic teaching or for any kind of teaching approach, such as the interactive whiteboard, are being used successfully for the dialogic approach (Mercer et al., 2010b; Major et al., 2018). Dialogue systems might seem more ideally suited for this approach, as they are based on dialogue, which is the core of dialogic teaching. There is in fact widespread interest in the use of dialogue systems in education, as exemplified by the EU-wide edubots project[5]. Dialogue systems can be designed specifically for pedagogical purposes, even for concrete courses (Sahil et al., 2016; Huang et al., 2019).

Dialogue systems are being used for students at different levels, from small children (Ruan et al., 2019) to graduate students (Huang et al., 2019); however, this technology seems to be used mostly for higher education (Kuyven et al., 2018). There is also a wide variety of subjects and skills that these tools are being employed for, but the main focus seems to be on STEM, with many systems also being used for linguistic skills development (ibid). Additionally, a common use of dialogue systems in educational contexts is not directly related to education: many systems are designed for tasks such as informing students about enrolment or exam dates (Sahil et al., 2016).

As we have mentioned, some attempts have been made to use dialogue systems to develop linguistic skills - the focus is primarily on linguistic skills in a foreign language (Goda et al., 2014; Ruan et al., 2019). It can thus be said that systems exist which can help students reinforce at least some oracy subskills. One notable example is the work done by Catania et al. (2020). They provide a framework, tested through a Wizard-of-Oz study[6], for oracy skills development in children's native Italian language. The task to be completed

[5]Link to Edubots project

[6]In a Wizard-of-Oz study, subjects interact with what they believe is a fully automated system, but there is a person secretly controlling the system output. This allows researchers to obtain semi-real usage data before a system can be developed with that data or the insights extracted from it.

through their system consists on describing their physical appearance to create an avatar; the focus is thus on linguistic subskills, though other subskills may also be practiced to some extent (e.g. the social subskill of responding appropriately or the physical subskill of speaking in a clear voice) (Mercer et al., 2017). Perhaps part of the potential shown by this system stems from the fact that it was a task-oriented system, which makes the interaction easier to control (though the Wizard-of-Oz study makes it impossible to confirm whether this controlled setting would result in robust natural language understanding). Given the simplicity of the task, it cannot be said to meet all the principles of dialogic teaching (Alexander, 2010), but it is certainly purposeful and supportive. Many dialogue systems used in education are designed as chatbots (not geared towards completion of a task) using AIML[7] (Kuyven et al., 2018). This language makes it relatively easy to design a dialogue system (Heller, 2016; Kuyven et al., 2018). The resulting chatbot, however, may not converse with enough coherence for oracy skills development (Goda et al., 2014).

3 Review of technical literature

3.1 What are dialogue systems?

As explained in the Definitions section (1.2), dialogue systems are programs that "communicate with users in natural language" (Jurafsky and Martin, 2019). These programs can be task-oriented systems, which serve to complete a task, or chatbots, which simply converse to entertain users or improve their experience with a task-oriented system (Chen et al., 2018; Jurafsky and Martin, 2019). As is the case with other language technologies, there are many approaches to building dialogue systems, depending on their function and the resources available.

The earliest system, chatbot ELIZA, from 1966, was built using rules (Jurafsky and Martin, 2019). These rules matched word patterns in the user's input and returned output with a new pattern including selected elements of the input pattern and, when necessary, transformations of the input (e.g. a word like "my" in the input would become "your" in the output) (ibid). These rules also stored any input that was considered important, in order to refer back to it when new input matched relevant keywords from the memory (ibid). This architecture is still used nowadays, especially after ALICE, essentially a more complex version of ELIZA, was released, introducing the now very popular AIML (Artificial Intelligence Markup Language) (Heller, 2016). AIML is widely employed in educational chatbots, possibly due to its ease of use (Kuyven et al., 2018).

An alternative to the time-consuming task of writing rules is using corpora to build the dialogue system. Dialogue corpora can be used to retrieve output directly, comparing the user's input to utterances in a corpus and either returning the answer to the most similar utterance, or returning the corpus response to the most similar utterance; the latter

[7] Artificial Intelligence Markup Language. This markup language uses pattern matching to link user input to suitable prewritten response templates. Templates can also include patterns to adapt to the context of the conversation.

seems to yield better results (Jurafsky and Martin, 2019). Another option is using corpora to train generative, instead of retrieving responses (Jurafsky and Martin, 2019). Neural approaches offer an effective alternative and are gaining popularity (Chen et al., 2018). One advantage of neural systems is that they can return output not found in the training corpus; one disadvantage, on the other hand, is that they will normally return generic, "safe" responses (Chen et al., 2018; Jurafsky and Martin, 2019). This can be remedied by using reinforcement learning introducing variables that account for response diversity or by using adversarial networks (ibid). Though time-saving end-to-end neural systems have been developed, these require large amounts of data (Chen et al., 2018); such data may not be easily available for many languages and domains. Systems are more commonly developed using a pipeline architecture, a version of the 1977 GUS architecture, which may combine rules and neural networks (Chen et al., 2018; Jurafsky and Martin, 2019). In these pipelines there are four key components: a Natural Language Understanding (NLU) module, a Dialogue State tracker, a set of dialogue policies, and a Natural Language Generation (NLG) module (ibid).

The NLU module processes the user's natural-language input to perform intent detection and slot filling. An intent is what the user attempts by inputting something. For instance, "I want to book a table at Giovanni's" should be assigned an intent like "request - restaurant booking", and the completion of the task would require filling slots with entities about the time and date, number of guests and restaurant name. Slots store entities that are important for the conversation, especially when a task is to be completed, and which slots may or must be filled depends on the intent that is assigned to the utterance (ibid). For example, continuing with the restaurant example, as the intent was "request - restaurant booking", the system would not try to fill irrelevant slots like "song to be played" or "dress size". The dialogue state tracker is a subsystem that analyzes the current and previous utterance, as well as the filled slots and intent; the information from the state tracker is then used to apply the right policy (ibid). In the previous example, if the dialogue state tracker observes that the intent is booking a table at a restaurant and the slot "restaurant name" has been filled directly from the request, it may select a policy where the system asks information to fill the next missing slot (e.g. date, time or number of guests), or it could also select a policy seeking confirmation if there are several restaurants called Giovanni's which could fill the slot (e.g. "Giovanni's in High Street or Giovanni's Pizza in Clark Street?"). Finally, the NLG creates the natural-language output that matches the selected policy (ibid).

3.2 Important design principles for dialogue systems

For the design of a dialogue system, other aspects must be borne in mind besides the overall architecture. People are used to the conventions of human dialogue, so these need to be taken into consideration for dialogue systems to have successful conversations with people. The most popular explanation of how humans communicate successfully are Grice's four maxims, which can be summarized as follows: we need to give the right amount of information (enough, but not too much), we need to give true information, we need to give information that is relevant to the conversation, and give our information in a clear, organized manner (Grice, 1975, in Jurafsky & Martin, 2019). These maxims form the cooperation principle, which posits that speakers cooperate to communicate, each trying to follow the maxims for communication to be effective (ibid). Another key idea of human conversation is that, by definition, dialogue involves more than one active participant, or else it would be a monologue (Mercer et al., 2019). A consequence of this is that participants will need to take turns to talk, and in human conversation these might overlap; spoken dialogue systems thus need to analyze prosodic cues like pauses, as well as other available cues, to detect when the user's turn is over and the system can intervene (Jurafsky and Martin, 2019). The presence of several participants also has as consequence that these need to take the initiative to start and continue the conversation; in most human conversation, this role alternates between participants (ibid). Mixed initiative is very difficult for dialogue systems to achieve, which is why the initiative is normally left to the user, to avoid the technical hurdles of mixed initiative, but also because systems constantly taking the initiative can sometimes lead to user frustration (ibid). Another important concept is the idea of implicature: people's utterances do not always contain all the information explicitly, as they can rely on other human interlocutors to disambiguate references to previous points in the conversation or to shared world schemata; that is why it is important for systems to be able to keep track of the conversation and the previously mentioned entities (ibid).

Given the complexity of human dialogue, it is to be expected that there might be miscommunications even with the most advanced systems - after all, miscommunication also happens between humans. It is thus important for systems to have policies that help them deal with errors successfully (Jokinen, 2009). A sophisticated error-handling strategy distinguishes between error types (no input detected, input could not be interpreted, or system unable to process user request) to select the most appropriate response, like a reprompt (repeating the system's question), an apology, making a suggestion, etc. (Google, 2021). What response is most appropriate also depends on the system's confidence that it interpreted the input correctly and how important the information from the input might be: for example, before making a purchase, the system needs to be perfectly confident that it will perform the exact purchase that the user requested (Jurafsky and Martin, 2019). How many errors have already been encountered in the conversation is another key variable; if the system's first response to an error does not help solve it, repeating the same approach would result in an endless cycle of frustration for the user (Google, 2021).

The success of the conversation does not only depend on the system making few or no errors and a goal being achieved; efficiency, or rather perceived efficiency is also crucial (Jokinen, 2009). How long users feel they wait for a system response has been found to be more important than actual waiting time; giving the user some form of feedback while their input is further processed can help them feel that the cooperation principle is being upheld (ibid). It is also important to consider that not all tasks or task components require natural language understanding or generation, and they may be carried out more efficiently through a direct manipulation interface (e.g. buttons) (ibid).

Lastly, we must also mention the media equation, a theory that claims that people tend to treat computers and other media as people (Reeves & Nass, 1996, in Heller, 2011). This means that, when people talk to a dialogue system, they associate it with a certain personality (e.g. intelligent, dim-witted, submissive, funny, friendly, etc.). This is especially problematic considering that systems tend to be given female voices and names, together with stereotypical personality traits (e.g. submissiveness), which may reinforce gender roles and inequalities (Jurafsky and Martin, 2019; West et al., 2019). Therefore, dialogue system development requires making informed, careful choices about which personality the system is intended to transmit through its utterances, voice or possible avatar. Persona design also allows developers to better control that users perceive the system in a way that is more conducive to successful communication (Google, 2021), especially if this involves clearly showing the user what the system's capabilities are and how they might interact (Marge et al., 2020).

3.3 Argumentative dialogue systems

In our review of the literature on dialogic teaching, we have mentioned discussion as the main kind of interaction needed, where different participants present and argue ideas and build upon each other's contributions. Then, would dialogue systems be able to carry out this type of talk? More research is still needed for dialogue systems to successfully participate in conversations with more than one interlocutor (Marge et al., 2020). Nonetheless, advances have been made which show that dialogue systems are capable of engaging in debates with one interlocutor. The pioneer of this type of system can be found in IBM's Project Debater (IBM, 2021). This system, given a debate topic, could scan billions of text lines and generate a four-minute opening statement defending a position; after listening to the opponent's statement, "she" could generate a counterargument and then a closing statement summarizing the main ideas of the debate (ibid). The system is even able to use resources that humans employ for persuasion, such as humor or appeals to emotion (ibid). Although the technical and human resources available at IBM would not be at everyone's reach, the data from the project has been made public, which could enable the creation of other debater systems. The field of argument mining provides further resources, like the ArguAna datasets [8], which have been demonstrated to be usable for a rudimentary debater system (Kulatska, 2019), or models for argument generation (Schiller et al., 2021, in Gurevych, 2021).

4 Linking dialogic teaching with dialogue system features

In the following subsections (sections 4.1 through 4.5), we present a framework that lays out the links we have established between aspects of dialogic teaching and features of a dialogue system which could assist in the implementation of this pedagogical approach. Each of the

[8] Link to ArguAna datasets

five sections focuses on one of the five principles of dialogic teaching (see section 2.1.1): cumulative, collective, purposeful, reciprocal, supportive. The first part of each section lists the aspects of dialogic teaching that we link to the features enumerated afterwards; these are the features that we consider a dialogue system would need in dialogic teaching. Our classification of the aspects of dialogic teaching was built around the principles described by Alexander (2010), which we link to the more easily measurable indicators described by Sedova (2017) (see section 2.1.3) and the oracy sub-skills detailed by Mercer et al. (2017) (see section 2.2.1). Though, in our literature review, we also explained the valuable contributions by Mercer et al. (2010b) in describing effective and ineffective classroom talk, we did not include any links to the different types of talk in our framework - after all, we would only say that exploratory talk would need to be promoted, and what that entails is described with more detail by the principles and indicators.

4.1 Feature set 1: Cumulative principle

- **Principle:** Cumulative
- **Indicator:** Uptake
- **Oracy skills:**
  - Cognitive
  - Social and emotional
  - Linguistic
- **Dialogue system features:**

1. Memory to store the main points of the conversation and build upon them; it will need to remember which arguments have already been presented and avoid repetition. Displaying chat logs so that the student can also remember what has been mentioned (Mercer et al., 2010b; Mercer and Howe, 2012; Major et al., 2018).

2. Generating questions on the topic of the discussion (Alexander, 2010), so it will need to identify the topic and have or gather information about it to make coherent contributions (Goda et al., 2014).

3. Stance detection to analyze the student's arguments and present counterarguments and respond coherently (Goda et al., 2014; Stab and Gurevych, 2014; Kulatska, 2019).

4. Encouraging the student to use their own words to ensure that they understand the topic (Mercer et al., 2019), but at the same time introducing appropriate terminology (Mercer and Howe, 2012).

4.2 Feature set 2: Collective principle

- **Principle:** Collective
- **Indicator:** Open discussion
- **Oracy skill:** Social and emotional
- **Dialogue system features:**

1. Measuring the length of the exchanges: neither the system nor the student can monopolize the conversation. Most exchanges should be somewhat lengthy to have substance (i.e. reasoning, building up on previous ideas...) (Sedova, 2017; Mercer et al., 2019).
2. Presenting a wide variety of ideas on the topic, for them to be evaluated and accepted or challenged; a student-dialogue system interaction could have a very limited perspective, which is why an effort should be made for the system to present more than one voice or point of view (Alexander, 2010; Mercer et al., 2019).
3. Matching the topic of the system's and the student's utterances: the speakers should not lead parallel conversations (Goda et al., 2014; Mercer et al., 2017, 2019).

4.3 Feature set 3: Purposeful principle

- **Principle:** Purposeful
- **Indicator:** Student reasoning
- **Oracy skill:** Cognitive
- **Dialogue system features:**

1. Setting the goal of the exchange: it should not be mere chit-chat but reach some kind of understanding on a topic (Alexander, 2010; Sedova, 2017).
2. Showing the student what can be achieved through the discussion; the task can be more successful if the students see value in it (Mercer and Howe, 2012; Thies et al., 2017).
3. Providing background information that can serve as the basis for the student's contribution (Kulatska, 2019).
4. Feedback and access to chat logs to help the student reflect on their progress, also to increase accountability and motivation (Mercer et al., 2010b; Major et al., 2018; Huang et al., 2019). Correcting the student on something that they did well but the system misinterpreted would cause frustration; however, if the students are not

provided feedback throughout the task, they may not engage with it as desired. A good compromise might be to give frequent feedback which acknowledges that the student's work is in progress, that in open discussions there is not absolute right or wrong, and that the dialogue system is not infallible (Pinkwart et al., 2008; Kulatska, 2019).

5. Identifying reasoning in the student's answers; it could be expressed implicitly (Jurafsky and Martin, 2019). When no reasoning is detected, the system should encourage reasoning (Mercer et al., 2010b; Mercer and Howe, 2012; Sedova, 2017). The system should also model productive dialogue by reasoning (Mercer et al., 2010b; Mercer and Howe, 2012; Okada et al., 2018).

4.4 Feature set 4: Reciprocal principle

- **Principle:** Reciprocal
- **Indicators:**
  - Student questions
  - Open discussion
  - High-order questions
- **Oracy skills:**
  - Cognitive
  - Social and emotional
- **Dialogue system features:**

1. Countering the student's arguments or reacting with questions to delve deeper into the topic and develop ideas (Alexander, 2010; Mercer et al., 2019).

2. Showing openness to questions by explicitly saying that it's open to questions, presenting information and telling the student to ask about it or presenting something new/strange/opposed to the student's view that may inevitably lead to questions (Alexander, 2010; Sedova, 2017).

3. Asking for clarification when the student's utterances are not understood (Jokinen, 2009; Bii et al., 2013). Giving explanations considering the level of understanding that the students have shown (Mercer and Howe, 2012).

4. Dialogue systems may rely on the user taking charge of the conversation (user-initiative systems) or instead have it completely controlled by the system (system-initiative systems). A more dialogic exchange might require the more complex and human-like mixed-initiative design, where either participant (student or system) may direct the conversation at different points (Jurafsky and Martin, 2019).

5. Helping the students see how much the dialogue system can understand. Users often address machines as they would people, so they need to be made aware that the dialogue system may understand less than a person, but also that they do not need to oversimplify their utterances. In the context of dialogic teaching, it might be especially important to show users that the dialogue system can process complex questions that go beyond fact regurgitation, so that students are encouraged to include more cognitively demanding questions (Jokinen, 2009; Thies et al., 2017; Huang et al., 2019).

6. Asking "high order questions", more linked to reflection than to information retrieval (Mercer et al., 2010a; Sedova, 2017). To discuss topics where many perspectives can be considered and questions have no single answer, the system would need very large and diverse data sources and a deep-learning approach to generate answers that seem natural and contribute to the discussion (Kuyven et al., 2018; Kulatska, 2019; Jurafsky and Martin, 2019).

4.5 Feature set 5: Supportive principle

- **Principle:** Supportive
- **Indicator:** [Not directly observable]
- **Oracy skills:**
  - Social and emotional
  - Linguistic
  - Physical
- **Dialogue system features:**

1. Language that, even when used to challenge students' ideas, is respectful and encouraging (Alexander, 2010; Mercer et al., 2010b; Thies et al., 2017; Major et al., 2018; Ruan et al., 2019). A friendly avatar could be helpful (Bii et al., 2013; Huang et al., 2019) if its design is not distracting (Heller, 2016).

2. Detecting confrontational talk and demanding that students express their ideas in a respectful, reasoned manner (Mercer et al., 2010b).

3. Multimodality, to reinforce physical and linguistic skills and to adapt to different learning styles and accessibility needs (Jokinen, 2009; Heller, 2016; Major et al., 2018). Multimodality involves voice recognition and speech generation but may also involve a static or dynamic avatar (Bii et al., 2013; Heller, 2016).

5 Designing an argumentative task

In order to translate the framework from section 4 into a specific dialogic task, we looked at studies that featured practical examples of how dialogic teaching may be conducted. As dialogic teaching is a pedagogical approach and not a specific methodology (Alexander, 2010), it can take many forms. The source we found most helpful for our task was the study carried out by Sedova (2017), which draws its conclusions from extensive classroom-observation data and provides examples of such data. The most successful of the examples provided (in that it showed the most indicators of dialogic teaching taking place) was a lesson where the class read a text and then debated the ethical implications of what they had read (ibid).

Based on the task described by Sedova (2017), we designed an argumentative task; we believe this type of task to be a good option due to the importance of discussions in dialogic teaching (Alexander, 2010) and the emphasis that this teaching approach puts on ideas being reasoned (Mercer et al., 2010b, 2017). The proposed task will be described in sections 5.1 through 5.8. However, for clarity, we include here a summary of the subtasks (ST) that the task will consist of:

1. Introduction and guidance
2. Text reading
3. Paragraph identification (thesis and arguments)
4. Thesis and argument rephrasing
5. Scaffolded discussion

Before describing the task, in section 5.1 we will first describe the target audience of the task. In section 5.2, we will justify our choice of reference material (SAT Essay texts). The description of the task and its subtask is given in section 5.3. Then, in section 5.4, we comment on some general design decisions with regard to the system persona, and in section 5.5 we briefly describe a proposed interface for the system. Section 5.6 then introduces Appendix A, where we list and illustrate the dialogue acts of our proposed system. Section 5.7 introduces the previous appendix, Appendix B, where we give an example of what form the task might take. Then, in section 5.8, we compare our proposed task with the framework that we detailed in section 4.

5.1 Target users

Taking the example by Sedova (2017) as the basis for our task, we set our target audience as high-school students, for three reasons, detailed below.

Firstly, studies on the use of dialogue systems for education focus mostly on higher education instead of the lower levels (Kuyven et al., 2018).

Secondly, oracy skills are very important for students' academic and work life (Mercer et al., 2017). In most countries, the duration of compulsory education corresponds to the years needed to complete secondary education [9][10], which means that, after high school, some students may enter the workforce and not complete any more years of education, while others may pursue some form of higher education. In both cases, oracy skills would be necessary (Mercer et al., 2017), especially when students join a competitive, globalized academic world or labor market (Okada et al., 2018). Thus, oracy skills need to be developed in the lower stages of education to benefit all students.

Thridly, designing a system aimed at a younger audience (i.e. children instead of teenagers) would involve greater challenges that might be unnecessary in a preliminary study like this. The main reason for this is that speech recognition performs more poorly with children speech (Yu et al.); this is partly due to the distinct characteristics of children's speech (ibid), which could also complicate the task of designing responses of the appropriate level.

We are also designing the task specifically with American English-speaking students in mind due to the higher availability of resources for that setting, though the task might be suitable for English speakers from other countries or non-native English speakers.

5.2 Use of SAT texts

Linking the target audience with the example tasks described by Sedova (2017) brings to mind college admission tests, many of which involve analyzing and/or developing an argument – some examples are the American SAT (CollegeBoard, 2015), the German Abitur[11] or the Spanish PAU[12]. As we have selected American high-school students for our target audience, we use the SAT as our basis for the task - as in the task described by Sedova (2017), a text is used to start a discussion. Below we further justify why we selected the SAT as our source for texts.

Texts as domain knowledge source. A task that can be performed by a wide variety of students may require a given text as its base, so that students with different background knowledge, exposed to different curricula and teaching methods, can have the same information needed to complete the task. Otherwise, if students using a dialogue system are just asked to discuss a topic, they may not know enough about the topic if it was not mentioned in their particular class. Therefore, the system will need to provide information. For example, Arguebot (Kulatska, 2019), one of the few existing dialogue systems that are similar to this project, provided a text with information on the topic of debate, even though the users could choose the topic; this information was considered useful by over 80% of testers (ibid). Those who did not find the text useful claimed that they would have preferred the information in a more streamlined format, such as an outline (ibid); however, it must be borne in mind that Arguebot was not designed for pedagogical

[9] Link to World Bank data
[10] Link to Wikipedia article on Compulsory education
[11] Link to description of Abitur Task 3b from the Baden Württemberg education portal
[12] Link to the description of the national college admission test at the Spanish legislative portal

purposes, and thus the improvements it might require may not apply to dialogue systems with a different purpose.

Texts as models. Providing a text as basis for the argumentative task also gives students a model to learn from. According to Observational Learning Theory, analyzing a model can be conducive to learning; this theory is considered relevant especially with skills related to language, as these require the student to practice, because those skills cannot be learned through mere observation (unlike what its name may suggest, Observational Learning Theory involves more than observation, observation is only the first step) (Okada et al., 2018). Thus, the student needs to first observe the model, but then also analyze it, imitate it, and finally reflect on their performance, which could be fostered by being given feedback (ibid). In the next section (5.3) we describe our proposed task: the first subtask being an introduction, subtask 2 would cover the observation of the model, subtasks 3 and 4 the scaffolded analysis, subtask 5 the imitation of the model, and feedback would be provided throughout the entire task.

Texts to reinforce dialogic principles and oracy skills. Including a text in the task also adds a voice aside from the student's and the dialogue system's. Ideally, a discussion should involve more than two participants (Alexander, 2010), but not enough progress has been made on dialogue systems that can interact with more than one user at the same time, distinguishing each person's contributions (Marge et al., 2020). However, even if the student is performing the task alone with the dialogue system, the inclusion of an author's voice contributes to some extent to the principles of reciprocity and collectiveness, which might otherwise be absent in a context of student-computer interaction. Additionally, this other voice could help students develop certain oracy subskills (Mercer et al., 2017) that might not be reinforced in a different kind of student-computer interaction. These oracy subskills are primarily the social subskills; a discussion where the text's author cannot answer limits the social aspect of dialogue, but it at least involves the student paying attention to someone else's views. Analyzing someone else's argumentative text could also reinforce the cognitive oracy subskills of evaluating and summarizing ideas and building on others' views. Also, as has been mentioned, a text can serve as a model, and this model could help the student enrich their pool of linguistic resources to strengthen their linguistic oracy subskills.

Texts from appropriate sources. Using the SAT texts as the basis for the task ensures that the texts are suitable for high-school students, challenging but not exceedingly difficult, and that the students can work on them without prior knowledge of the specific topic being discussed (CollegeBoard, 2015). SAT Essay texts are selected by CollegeBoard researchers so that they "argue a point", are "written for a broad audience" and "use logical reasoning and evidence to support claims", which makes them not only suitable for high-school students, but also specifically for the task of analyzing someone's argumentation[13]. The SAT exam is taken by millions of students – for example, in 2020 more than 2.1 million students took the SAT Essay test in the US (CollegeBoard, 2020); due to this large number of test takers, there is an abundance of resources which can be used to train the dialogue

[13] Link to CollegeBoard's description of the Essay part of the SAT

system, even if we only consider resources from the redesigned test – the redesigned version of the exam was introduced in 2016 to better align the test with students' and colleges' needs (CollegeBoard, 2015). The SAT Essay test does not require students to provide any arguments of their own, as the focus is on seeing the students' reading and analytical skills as well as general writing skills (ibid). For our task to follow the principles of dialogic teaching and help students incorporate ideas into their own understanding and develop their own arguments, we must thus design a task that is different from the SAT Essay task, even if it takes the same texts as its basis. This difference between the SAT Essay test and our own task, however, means that the available sample answers will not be useful to us, which is one of the reasons why we are developing a dataset of artificial answers. This decision is further justified in section 6.1.7.

5.3 Task description

As explained above in section 5.2, we took the SAT as our source for texts to start an argumentative task; however, we needed to modify the SAT task to align it with dialogic teaching and so that it could be carried out with a dialogue system of the characteristics described in section 4. The concrete realization of the task that we arrived at, divided into five subtasks (ST), is as follows:

1. **Introduction and guidance.** Firstly, the dialogue system introduces itself and provides the student some guidance on how to use it and perform the task. We consider this guidance necessary for the student to be fully aware of the system's capabilities and thus use it to its full potential (Jokinen, 2009; Thies et al., 2017; Huang et al., 2019), as well as to ensure that the task is performed in the way most conductive to learning (Pinkwart et al., 2008). The system could also mention why the task might be valuable to the students to increase their motivation (Mercer and Howe, 2012; Thies et al., 2017).

2. **Text reading.** The actual task begins with the student reading the assigned argumentative text. The student is asked to read it out loud at least once for a voice recognition system to confirm that the student completed the reading; we believe this to be necessary because this type of task cannot succeed if the student does not complete the reading (Sedova, 2017), and students with little motivation to complete a task may only put in the minimum effort unless they are encouraged to perform the task in a more productive way (Pinkwart et al., 2008).

3. **Paragraph identification (thesis and arguments).** Once the student has read the text, they are asked to identify the paragraph that contains the author's thesis; then, the main argument (or any argument if they are all given the same weight by the author). The identification is done by selecting the paragraph number, in order to avoid using Natural Language Processing where it serves no other purpose than complicating the system (Jokinen, 2009).

4. **Thesis and argument rephrasing.** The student is then told to put the information they just identified (thesis and one argument) into their own words to understand and evaluate the ideas (Skidmore, in Mercer at al., 2019). The system identifies whether the student's input is an actual attempt at an answer and, if so, it evaluates the answer on the three feedback aspects we are considering (completeness, paraphrasing, correctness, detailed in section 6.1.6) and returns feedback that can guide the student towards a better answer.

   This is the subtask (ST4) we focus on for our dataset compilation (section 6.1) and our experiment (6.2).

5. **Scaffolded discussion.** The system then asks the student to develop their own argument on the issue of the text. To make this a more scaffolded task, the system guides the student through several steps (choosing a stance, providing arguments to back it, responding to counterarguments) until the student develops a solid argumentation (to determine when this goal is reached, quantity or quality criteria would need to be established). As tests concerning this part of the system are beyond the scope of this project, this part of the system is presented only as an option whose feasibility and suitability are to be confirmed in future studies.

   - In the first step, the system would ask the student what their stance is on the topic under discussion. If the student were unsure, the system could provide links to sites such as IdebateLink to the Idebate portal, which give a quick overview of important arguments for and against a debate topic. At any point in the following parts of the task, the student could be allowed to signal that they have changed their mind, and their argumentation would restart from this first step.
   - The system would then ask the student to provide an argument to defend their position. Again, sites like Idebate could be presented to students who cannot think of arguments to support that position; if the student defended the same position as the author of the model text, the system could also point the student to paragraphs where arguments were identified.
   - The system would then search its knowledge base for arguments that could challenge what the student said. This knowledge base would need to contain a large number of arguments classified by topic and stance, and the aspect would also need to be identified to match it with the aspect of the student's argument (Gurevych, 2021). Some of IBM Debater's datasets, such as XArgMining[14], could prove useful due to their size and topic/subtopic/stance/quality labels, if they could be aligned with the topics covered in the task.
     – If the system found no counterargument, it could simply ask the student to provide an additional argument.

[14]Link to the XArgMining dataset description at arXiv.org

  - If the system found a counterargument, it would present it to the student and ask them to provide their rebuttal. If adequate data was available, the system could give the student resources to build their rebuttal in case they are unable on their own. In either case, after the student provided a rebuttal or failed to do so, the system could ask the student to continue providing arguments to strengthen their argumentation.

- The system could end this part of the task after a specific number of arguments and/or rebuttals were reached – the specific number would have to be decided based on the data available, the topic and the level of the students. If argument quality were measured, the task could be ended after a specific quality score were reached.
- The student would then be asked to place their arguments and rebuttals of the system's counterarguments in a diagram, as these tools have proven useful for students' argumentation (Pinkwart et al., 2008). The student could then use the diagram to help them participate in class discussion, write an argumentative essay or simply as prove of task completion or as a way to reflect on their performance.

This task has been thought of as a scaffolded task, with a focus on guiding students to help them improve their skills. For that reason, feedback is to be given throughout the whole task as needed, and the student would need to be able to access resources to help with the task. Feedback thus takes a formative role, rather than summative[15]. Nonetheless, tests with students and teachers would have to be carried out to ascertain whether additional feedback upon task completion might be useful.

5.4 System persona

It is necessary to design a personality for the system: users will inevitably project one as a consequence of the Media equation theory, by which people tend to treat computers as people (Google, 2021; Thies et al., 2017), or they may ask questions to learn about the system's personality instead of focusing on the task (Bii et al., 2013). If the persona is well designed, its effect on the user will be more easily controlled (Google, 2021). Due to the scope of this project, not much time can be devoted to this aspect. Tentatively, following the steps suggested by Google's guidelines and the insights from the literature on how young users interact with dialogue systems, the persona could be as described below, putting special emphasis on the general characteristics and not so much on the details such as the name, visual representation and specific voice type.

Name: Robosan (tentative gender-neutral name pending change after data on user preferences can be analyzed).

[15]Formative feedback is given throughout the learning process to guide it, whereas summative feedback evaluates what students have learned at the end of the process, such as at the end of a course, semester or module (Douglas Brown, 2004).

Characteristics: friendly, not trying too hard to be funny, non-judgmental but not overly supportive (Thies et al., 2017); empathetic (Goda et al., 2014; Thies et al., 2017) – using varied prompts can make the system sound less "robotic" and reinforce this illusion of empathy (Google, 2021).

Visual representation and voice: pending definition after data on user preferences can be analyzed

5.5 Interface

As can be seen in more detail in the example conversation (Appendix B), the suggested system interface is meant to contain more than only a chat window - the example only shows the chat, but other windows are referenced. As the task is based on a text, we suggest a separate window where the student could check the text more comfortably.

We have previously mentioned that diagrams have been shown to be useful tools for argumentation (Pinkwart et al., 2008). For that reason, we suggest the inclusion of a diagram tool that the students can use to organize the ideas that they come up with while speaking with the dialogue system. Ideas on how to design this diagram tool could be taken from already existing tools like LARGO (ibid), which would need to be adapted to a younger and less specialized target audience.

In order to keep track of student progress, a user authentication component is also suggested. This might help students feel more accountable (Major et al., 2018) and it might also ease teachers' monitoring of student work.

We also propose including microphone controls in the interface, so the students can signal when they want to start their turn and be heard by the system. This proved to be a useful feature in the Wizard-of-Oz study carried out by Catania et al. (2020). The students in that study were children with no experience using dialogue system, which the authors attribute to these systems not being widely used in Italy, where the study was carried out. Our target students may be expected to have higher skills and familiarity with dialogue systems (ibid). Still, given that our proposed system is meant to scaffold students so that they can develop their oracy skills, microphone controls could also be a positive feature, as this could bring their attention to turn-taking, an aspect of the social and emotional oracy skills (Mercer et al., 2017) (see section 2.2.1).

Adaptations to increase accessibility, though encouraged for everyone's benefit (Jokinen, 2009), are beyond the scope of this project and we thus offer no suggestions in this respect.

5.6 Dialogue acts

The tables included as Appendix A break down the potential interaction between the student and the dialogue system into all the dialogue acts considered. Jurafsky and Martin (2019) define a dialogue act as the function that an utterance performs in the dialogue, whether it is user input or a response from the system. For instance, one of the dialogue acts that we include in Appendix A, one performed by the system, is "Asking the student to present an argument", which could be realized in an utterance such as "Now tell me, which argument can you think of to defend your position?" Dialogue acts need to be correctly identified for the system to apply the most appropriate policy – deciding what the system needs to do (ibid). For example, the proposed system explains what the student is expected to do, and then it needs to correctly assess whether the student's response (e.g. something like "Wait, can you explain that again") is a request for clarification or a confirmation of comprehension - if it is a request for clarification, the system will need to provide an explanation, and if it is a confirmation of comprehension, the system will continue with the task.

Some of the dialogue acts can be performed without Natural Language Processing, by using a direct manipulation interface: having prewritten options that the user can choose from by clicking on a button or selecting option from a drop-down menu. This reduces the burden on the NLU module, improving the system's efficiency (Jokinen, 2009). The first example on Appendix A is an instance of this: "Okay, seems easy enough", an instance of the dialogue act "acknowledging comprehension". There are many ways in which a student could express that they understood the task instructions, but when we use a direct manipulation interface the system does not need to learn to recognize all those different forms that the dialogue act could take. This approach is only proposed for dialogue acts that do not require providing feedback on the student's utterance (e.g. when the student perform the dialogue act of "presenting an argument", it is important that they are able to use their own words, as this is part of how their oracy skills are assessed and developed).

5.7 Example conversation

Appendix B is an example of how a conversation between the student and the dialogue system might be once the system is finally developed with sufficient robustness. The example is intended only to make the task easier to imagine before a full set of prompts can be written and the system developed. This example does not contain all the types of input and output that could be expected and are contemplated in the dialogue acts table (Appendix A), but it aims to at least showcase a "happy path" that the conversation could take. As an ideal example, it presupposes that the system has robust NLU and sufficient data to return adequate responses, be it through retrieval or generation. Nonetheless, as mentioned in section 5.6, some dialogue acts are presented as elements of a direct manipulation interface to reduce NLU requirements. This example shows some possible features of the interface, though it should by no means be considered a final design, as tests are needed to select the most appropriate features. As mentioned in section 5.5, we propose including a microphone control, which in the example is shown as a button with a microphone icon. The system prompts were written taking into consideration conversation guidelines by Google (2021), as well as insights gathered from the literature. Nonetheless, of course, a Wizard of Oz study would be necessary to test students' response to the prompts and make whichever adjustments were necessary. In this example, the greeting omits the common step of asking the student their name – this information could be gathered by having the student log into the application, as suggested in section 5.5. We have opted for this method to increase student accountability and keep track of application use, which could reinforce this accountability and also provide useful information to teachers; eventual tests may shed light on whether this decision achieves the desired results or whether another approach is needed.

5.8 How the concrete task is linked to the general framework

Here we link the ideal features presented in our framework (section 4) to features of our example system. As we did in sections 4.1 through 4.5, we are grouping features by the main principle of dialogic teaching (Alexander, 2010) that they would enforce. Each principle is presented together with the associated indicators and oracy skills. Below each set of principles, indicators and oracy skills, we comment on the system features related to those elements of pedagogical theory. Each feature from our framework is presented in a bullet point, and below it we specify to which extent that feature would be present in our proposed task. It must be borne in mind that, to improve readability, the content from section 4 has been summarized and the author references removed. Throughout this section we mention by number the proposed subtasks described in section 5.3; for clarity, we summarize them below.

Task summary:

1. Introduction and guidance
2. Text reading
3. Paragraph identification (thesis and arguments)
4. Thesis and argument rephrasing
5. Scaffolded discussion

5.8.1 Feature set 1: Cumulative principle

Principle: Cumulative.
Indicator: Uptake.
Oracy skills: Cognitive, Social and emotional, Linguistic.

- Keeping track of which arguments have already been discussed.
  - To be implemented for ST5.
- Displaying chat logs.
  - To be implemented, allowing the student to scroll through the whole conversation and possibly to save it into a file.
- Generating questions on the topic of the discussion: identifying the topic and retrieving information on the topic

- Open questions not contemplated, only when asking the student to define their position on the topic.

  The text used in ST2 through ST4 allows the system to set and thus know the topic; datasets would be needed for the system to learn to identify subtopics in the students' arguments in ST5 and to retrieve relevant counterarguments.

- Stance detection
  - To be implemented for ST5.
- Encouraging the student to use their own words.
  - To be implemented in ST4.
- Introducing appropriate terminology.
  - To be presented by the text used in ST2 through ST4. Links to Wikipedia or other reference sites could be included to help students understand the terminology.

5.8.2 Feature set 2: Collective principle

Principle: Collective.
Indicator: Open discussion.
Oracy skill: Social and emotional.

- Measuring the length of the exchanges.
  - To be implemented in ST2 (to see if the student read the whole text), ST4 (to see if the student focused on the text parts relevant to the current stage of the task or if they instead summarized the whole text), and possibly in ST5 if the system had trouble analyzing lengthy arguments.
- Presenting a wide variety of ideas on the topic.
  - To be implemented to some extent in ST5 by introducing counterarguments.
- Matching the topic of the system's and the student's utterances.
  - The topic would be set by the text used in ST2 through ST4, allowing the system not to stray from it; the input from the student would be analyzed to detect when it diverged too much from the topic, and the student would be encouraged to return to the topic.

5.8.3 Feature set 3: Purposeful principle

Principle: Purposeful.
Indicator: Student reasoning.
Oracy skill: Cognitive.

- Setting the goal of the exchange and showing the student what can be achieved through the discussion.
  - To be implemented in ST1, where the system would tell the student the teaching objectives of the task, as well as its more tangible products (a diagram that could be used to develop a coherent argumentation).
- Providing background information.
  - To be implemented in ST5 when the student could not answer; sites such as Idebate[16] could be used to obtain this information.
- Access to chat logs.
  - To be implemented, allowing the student to scroll through the whole conversation and possibly to save it unto a file.
- Frequent feedback that motivates students and informs them of their progress.
  - To be implemented in all subtasks that require student input (ST2 through ST5).
- Identifying reasoning in the student's answers.
  - To be implemented in ST4 by analyzing whether the student included at least one argument from the text, and in ST5 possibly by matching the students' arguments with a larger dataset of arguments on the topic.
- Encouraging reasoning.
  - To be implemented in task descriptions (overall description in ST1 and specifically in ST5) and by asking the student to identify and rephrase arguments in ST3 and ST4 and to use their own arguments in ST5.
- Modelling productive dialogue.
  - To be implemented to some extent by using a model text in ST2 through ST4; however, the text would be rather monologic.

[16] Link to the Idebate portal

5.8.4 Feature set 4: Reciprocal principle

Principle: Reciprocal.
Indicator: Student questions, Open discussion, High-order questions.
Oracy skill: Cognitive, Social and emotional.

- Countering the student's arguments or reacting with questions.
  - Open questions not contemplated, only when asking the student to define their position on the topic. However, counterargumentation would be implemented in ST5.
- Showing openness to questions.
  - To be implemented to some extent in ST1, where the system should inform the student that they can ask for assistance; other types of questions are not contemplated.
- Asking for clarification.
  - To be implemented through the whole task: the system needs to be confident that it understood the student's input to provide adequate feedback – what level of understanding would be necessary would need to be assessed experimentally.
- Giving explanations considering the level of understanding that the students have shown.
  - To be implemented to some extent: the texts used as model for ST2 through ST4 were selected to be suitable for high-school students; resources for output generation in part 5 would also have to be adapted to the expected level of the students. Nonetheless, adaptation for individual level differences are not contemplated.
- Mixed-initiative design.
  - Design is system-initiative, as it is the dialogue system which guides the task. A mixed-initiative design would be exceedingly difficult to implement (Jurafsky and Martin, 2019) and the effort might not be warranted in a task more focused on scaffolding the student through a discussion than on having already skilled students lead an advanced discussion. An option which gives students control of their turns while still using a system-initiative system is the one used by Catania et al. (2020), where students had to press a button to mark the beginning and end of their turns. This also reduces the burden on the speech recognition system, as it would otherwise need to be adjusted to recognize utterances by young users, who tend to take longer to start their turns (ibid). This feature was explained in more detail in section 5.5.

- Helping the students see how much the dialogue system can understand.
  - To be implemented to some extent in ST1, where the system would give the student some hints about how they can communicate with the system.
- Asking "high order" questions.
  - To be implemented to some extent in ST5, where the student is asked to defend a position and then to respond to counterarguments.
- Large and diverse data sources and a deep-learning approach to generate answers that seem natural and contribute to the discussion.
  - To be implemented at least to some extent in ST5: diverse datasets would be needed to return output based on either information retrieval or on deep learning – which design is selected would have to be decided through tests.

5.8.5 Feature set 5: Supportive principle

Principle: Supportive.
Indicator: [Not directly observable].
Oracy skill: Social and emotional, Linguistic, Physical.

- Respectful and encouraging language.
  - To be implemented throughout the task: prewritten answers are to feature a respectful and encouraging style, and data used for output generation or retrieval would have to be reviewed not to include disrespectful language.
- Friendly avatar.
  - The literature on avatar design was beyond the scope of our review; we can thus offer no suggestions on the concrete realization of this feature.
- Detecting confrontational talk and demanding that students express their ideas in a respectful, reasoned manner.
  - To be implemented to some extent: in ST2 through ST5, students are asked to accompany claims with arguments. Our planned experiments contemplate disrespectful language in off-task student answers; further tests would have to be performed to detect such language in otherwise correct answers.
- Multimodality (voice recognition, speech generation, a static or dynamic avatar).
  - To be implemented to some extent: though our experiments are text-only, the task is designed to feature speech recognition and generation. As mentioned above, no suggestions are offered regarding avatar design.

6 Assessing task feasibility: data and testing

As can be gathered from section 5, the proposed dialogue system would be a complex tool consisting of several subsystems to control the numerous subtasks. It would also require large amounts of data to be trained to analyze students' utterances, as well as to retrieve or generate counterarguments and point to the most relevant information sources. Given the complexity of the system and the small scale of our project, we have decided to start our tests focusing on ST3 and ST4, where the student selects the paragraphs with the thesis and an argument, and where the student summarizes the author's thesis and main argument. Our tests involve compiling a dataset with, first, annotated texts where the student could be asked to select the paragraphs with the thesis and an argument (ST3), and, secondly, possible answers that the students could give when prompted to summarize the thesis and argument (ST4). The dataset is described in detail in sections 6.1.1 through 6.1.7. The tests also involve carrying out an experiment (sections 6.2 through 6.2.5) using the dataset to determine whether a selected tool (mentioned in section 6.2) could help the system analyze students' answers to ST4 in order to provide adequate feedback.

It might seem more logical to start from the beginning, ST1, but since that part is concerned with guiding the student to complete the task, it might be better to tackle this aspect after having at least a first version of the rest of the system. It might then seem that ST2 is the second most logical place to start. However, experiments regarding this part would require a voice dataset, which would require access to a sufficiently large pool of subjects – especially given the idiosyncrasy of speech recognition with young subjects (Yu et al.). For this same reason, we are performing our experiments using only text input and output, without adding voice. The next subtasks, ST3 and ST4, can be studied more easily with artificial [17] written data; we believe that this imperfect data might suffice for developing a first sketch of these system components. Further experiments with real data may prove or disprove the usefulness of these initial efforts – at any rate, we hope that we can provide the structure for the ulterior collection of real data.

6.1 Dataset compilation

In sections 6.1.1 through 6.1.8, we describe why and how we developed a dataset to analyze what resources might be necessary for ST3 and ST4, the subtasks that we have decided to focus on for the reasons mentioned in the previous section (6). We divide this section into eight subsections. In section 6.1.1, we explain the need for a dataset and we justify why we use some already available data and add some new artificial data. In section 6.1.2, we describe the argumentation model that we use to annotate the structure of the texts that the student would need to analyze in ST3 and ST4. In section 6.1.3, we justify the textual unit we use for our annotation: the paragraph. In section 6.1.4, we discuss the possible ways in which the annotation labels can be assigned to the paragraphs - the ideal case

[17] We refer to this data as artificial because, though it was written by a person, it did not come from the target users of the system nor was it written under the target conditions of completing an educational task.

of "one unit matches one label" and some exceptional cases where a paragraph requires more than one label. In section 6.1.5, we explain a type of content we decided to add to our annotations to expedite the process and as some possibly valuable information for the system: summaries of each paragraph function in a text. In section 6.1.6, we move on to the criteria used to annotate the data of student answers as good or bad. In section 6.1.7, we describe how we develop answers data to match those criteria and in which proportion. Finally, in section 6.1.8, we briefly discuss the size of the dataset that we were able to compile.

6.1.1 Dataset justification

Dialogue systems can generally be described as functioning based on rules or corpus data (Jurafsky and Martin, 2019). Some rule-based systems can function adequately, such as the popular ELIZA chatbot (Heller, 2016; Jurafsky and Martin, 2019), but without corpus data a chatbot's capabilities are very limited[18]. Even if the architecture relies solely on hand-written rules, the people writing the rules will need some data to use as reference. Thus, even for a small project like ours, carrying out preliminary tests for a component of a simplified system, data is essential. We need data in the form of texts that the students would analyze (ideally a dataset of texts adequate for students, with their argumentative structure annotated in a simple but informative manner), and as sample answers that the students would give when prompted to analyze the texts (ideally brief yet complete theses dealing with the same precise topics as the texts from the text dataset).

As mentioned in section 5.2, we selected the SAT test as a good starting point for our task – SAT tests provide us with appropriate texts that can serve as basis for the task that the system is to perform with the students. However, the SAT Essay requires students to write a complete essay analyzing all the resources that the authors uses in their argumentation, but never discussing their own opinion (CollegeBoard, 2015). This means that SAT Essay answers would certainly not be useful for ST5, where students are asked to develop their own argumentation. The SAT Essay answers are also not useful to train the system to evaluate student's answers to ST3 and ST4 for two reasons:

Structural differences: In the SAT Essay test, the student answers by writing their analysis as an essay. The purpose of our task is not merely evaluative, but primarily formative, which means that the task is broken down into simpler tasks, providing the **Availability:** Even though test prompts are readily available, the same is not true for test answers, whether scored or unscored.

We looked for suitable data in several datasets, but rejected them. We considered datasets by IBM[19], the args.me corpus[20], datasets from ArguAna[21] and the Ubiqutous Knowledge Processing Lab[22]. One reason why we rejected them was the complexity of

[18] ELIZA's capabilities can be tested on numerous websites, such as this link
[19] Link to the IBM Project Debater datasets
[20] Link to the args.me dataset at Zenodo
[21] Link to the Arguana datasets
[22] Link to UKPL resources at Darmstadt Technical University

the annotations, not suitable for comparisons with analyses to be made by high-school students. Another reason was that they contained very short texts or texts of inadequate quality to serve as educational models or as basis for an analysis. Some datasets were also rejected because they contained isolated arguments or claims, not complete argumentative texts - these could however be useful for ST5, where the system is to analyze the students' own arguments and respond with suitable counterarguments. Despite the limitations of using only the SAT texts, it was considered the best option, if only as basis for the task.

Having selected the text data that the students would analyze, the next challenge was obtaining sample answers that they could give. As mentioned above, the datasets that we found were not suitable. A Wizard-of-Oz experiment would be the ideal means of obtaining a dataset to train the system (Thies et al., 2017). In this ideal setting, a large and diverse sample of target students would complete the task on a computer, believing they are talking to a dialogue system, but it would be a person, trained for the task, who would be talking to them. This would return data that could be considered real data, as it would have been produced by target users in a context made to simulate the target setting as closely as possible. However, such an experiment would be costly. Large, high-quality public datasets would have been another good alternative. As we mentioned, we looked at argument-mining datasets to see if they could be used to train our system. However, given the specificity of our task, available datasets did not seem useful: some were only annotated for topic but not stance, the ones annotated for quality followed criteria not aligned with ours, the topics covered did not match the SAT tests, and none were obtained in the context of high-school students completing a task. For these reasons, it seemed more feasible to perform our tests with an artificial dataset created by ourselves. Writing our own answers would allow us to control the variables that we wanted the system to analyze, both in terms of type and sample size. Naturally, artificial answers cannot be expected to perfectly resemble real data, but they can serve as a starting point to develop the basis of a system which can then be tested on real users to obtain real data that can be used to further develop the system (Jurafsky and Martin, 2019). Additionally, despite the limitations on the answers part of the dataset, the text data is from the source we believe to be most appropriate, as justified in section 5.2. Thus, we provide a solid base that might be of sufficient quality for future data gathering through Wizard-of-Oz studies.

6.1.2 Argumentation model

Naturally, the analysis of argumentative texts requires a model to guide it. One of the most influential models is probably Toulmin's (Toulmin, 2003, in Andrews, 2005). Andrews places it around the middle of a spectrum that ranges between formal logic and rhetoric. This middle point is believed to be appropriate for education, where argumentation occurs in different disciplines and contexts (ibid). Argumentation is described as a process linked to "transformation, clarifying and changing ideas, personal growth, identity formation, and other dynamic aspects of learning", not bound by the conventions of a single textual genre (ibid, p. 110); especially this last point makes it necessary to adopt a flexible model, far from the rigidity of formal logic, but with enough structure to be used by a computer

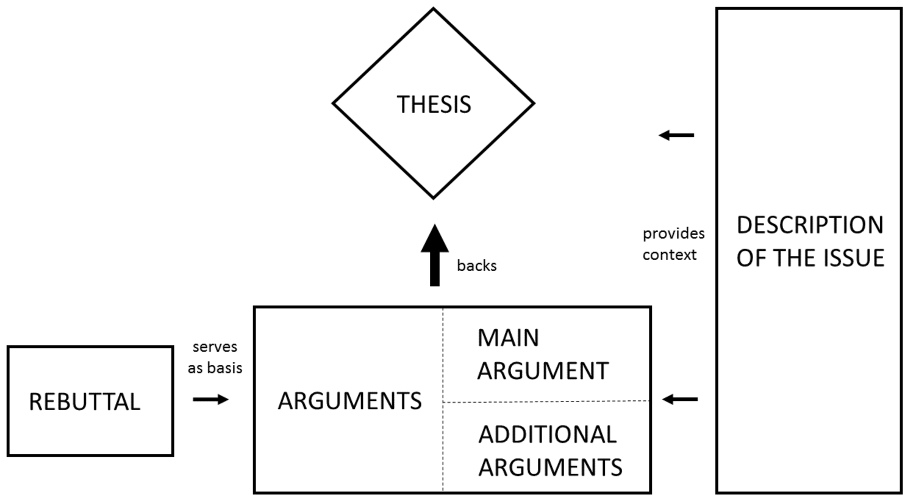


Figure 1: Argumentation model used to annotate our SAT texts

system. In Toulmin's model, the main elements of an argument are claims (defending a position), grounds (evidence to justify defending that position), warrants (linking the evidence to the claims), backings (justifying the warrants) and possibly qualifiers (modifying the link between grounds and claims) and rebuttals (challenging the relationship between grounds and claims) (Andrews, 2005; Stab and Gurevych, 2014).

It seems necessary to arrive at a highly simplified model, due to the level of the target audience of the dialogue system (high-school students), the limited capabilities of a simple dialogue system and the difficulty of annotating arguments coherently (Gurevych, 2021). Stab and Gurevych (2014) provide a useful example of such a model in their guidelines for the Argument Annotated Essays Corpus (AAEC) (ibid); we take their model as an important reference because it is intended for use in Computational Linguistics. They later describe a more complex structure, which is justifiable given the target users of their dataset (computational linguists). However, their description of the elements of an argument begins with only claims and premises to back the claims. They later establish a hierarchy, distinguishing between claims and major claims. In our annotation scheme, we thus establish as the main elements the thesis (the major claim) and the arguments that support it (the premises). This structure may suffice to classify all the SAT texts we encounter, though some texts and text chunks require departing from this simplest of models and adding some more elements, while attempting to not complicate the model beyond what we consider essential. Figure 1 shows the final form of our selected model; below we explain the modifications we made to the initial simple model to arrive at the model shown in the figure.

One element that we have deemed necessary to add is the one we have labeled "description of the issue". Argumentation models, at least when intended for argument analysis (production lies beyond the scope of our experiments), are models for "distilling the salient from the residual" (i.e. separating the elements that contribute to an argument from other text parts that do not serve that purpose) (Andrews, 2005, p. 114). In order to provide the best feedback possible, we think that we also need to label these "residual" elements (at least as such, perhaps without analyzing them any further), so that if a student focused on the residual rather than on the salient we may inform them of the nature of their error. Aiming for the utmost simplicity, we have labeled the "residual" elements (i.e. not directly involved in the argument) as "description of the issue", since we have found that they generally serve the purpose of setting the context for the reader; it may be argued that they serve additional purposes, such as appealing to the reader's emotions with personal examples, but we believe that the "description of the issue" label encompasses most "residual" content in the simplest, broadest manner, including the examples. Still, we have reserved an "other" label for "residual" elements that do not fit the previous label (e.g. we have found a text where one paragraph tells the reader how to stay informed and engaged with the issue).

As we have described, the model by Stab and Gurevych (2014) includes hierarchical elements. While we tried to avoid complicating our model this way, it seemed justified to establish a hierarchy in some texts. We have observed some texts where the author seems to emphasize one argument over others; in such cases, we believe that a true understanding of the text on the part of the student would require seeing this distinction. We have thus decided to distinguish between main arguments and additional arguments in our annotation, but only where the author gives more importance to a particular argument, by mentioning it alongside the thesis and devoting more paragraphs to it than to other arguments. Another reason to justify this decision is that the "additional arguments" might not even be considered arguments in a more complex model, such as Toulmin's full model (Andrews, 2005; Stab and Gurevych, 2014), where they might be warrants linking the grounds to the claims. For instance, the May 2018 SAT text claims that research companies should make their data accessible to others; the main reason given for this is that medical research requires having all the data. One of the additional arguments given is that many of the studies that are made accessible do not include their data; in our simple model this has been labeled as "additional argument", but in a more complex model this could be seen as a warrant linking the need to publish data (claim) with the need for data in medical research (premise). Another example is the October 2018 SAT text about dwindling bat populations, which mentions both at the beginning and the end that bats are important to us because they eat insects. Thus, the argument about the value of bats to humans can be considered the main argument, while other arguments are mentioned once at different points in the text, but not emphasized in the introduction or conclusion (e.g. the argument about it not being costly for energy companies to make adjustments that would prevent bat deaths or the argument that, while curing diseases that affect bats is difficult, helping bats by preventing other causes of death is within our hands).

It must be noted that this distinction between main an additional arguments is based solely on the importance given to them by the author, not in their effectiveness. In ST3 and ST4, the student is not meant to evaluate the effectiveness of the arguments, only to identify them. This may reduce the learning gains from these subtasks, but this may be compensated by the following subtask (ST5), where the student does need to use critical thinking to evaluate their stance on an issue, present and support arguments and respond to counterarguments. At any rate, it must be born in mind that this is meant to be a scaffolded activity; only after students have developed their skills further can the scaffold be removed and more demanding tasks be presented to the students.

One last element that we needed to add to our simple model is the "rebuttal" label; this is one of the elements of Toulmin's full model (Andrews, 2005; Stab and Gurevych, 2014). In the earlier, simpler version of our model, we could consider any rebuttals as part of the argument that uses them as basis to challenge a view opposed to the author's. For example, the text from the December 2016 SAT mentions how some opposed giving workers unpaid leave for family reasons claiming that it would be bad for businesses; the author then quotes statistics on how this did not harm businesses, as a way to support their argument in favor of a new law making that unpaid leave paid. A complex argumentation model could separate the argument where some people believed unpaid leave to be bad for businesses from the argument where the author claims this turned out to be false; in our simplified model, we opted for considering both elements as part of the same idea that giving workers certain rights may not harm businesses in the end. Nonetheless, given that we established the paragraph as our working unit (more on this decision in section 6.1.3), it became necessary to add the "rebuttal" label for cases where the rebuttal and the argument ensuing from it are in separate paragraphs. For example, this happens in the May 2019 text (included as C). There, the author speaks in favor of a law requiring firms to disclose possible gender pay gaps. The author chooses to use a structure where they first present any counterarguments (e.g. paragraph seven covers most of them, saying that some believe the measure to be costly, complex and ineffective), to then contrast them with their arguments in favor of the measure in the following paragraphs (e.g. paragraph eight gives an example of a company where the measure worked well). This label could also prove useful in the unlikely but theoretically possible case where an author mentions some views opposed to the one they are defending and does not challenge these rebuttals (e.g. though an unchallenged rebuttal might weaken an argument, readers might appreciate an author acknowledging the limitations of their claims, especially when the author explicitly defends the "lesser evil", i.e. the less limited claim). Whichever the case, it must be noted that this label is only added for the system to be able to give adequate feedback: for example, if the student points to a paragraph containing only a rebuttal thinking that it is a supporting argument, this label can help the system explain to the student the nature of their error. That would be the only use for this label, as it is not part of what the student would need to analyze in the task: the task is concerned with the student identifying the key ideas of the text (thesis + arguments), not analyzing its entire structure.

6.1.3 Annotation unit

We have chosen the paragraph as our unit for annotation. We justify this decision with four reasons: avoiding conflicting views about how the texts are structured, adapting to the students' level, saving time on the annotation process, and respecting the author's choices. These reasons are further explained below.

Firstly, opting for larger units reduces disagreements about where elements of an argument begin and end; what matters in our system is simply where they are contained, not their precise beginning and end. The mere fact that Stab and Gurevych (2014) had to go into great detail in their annotation guidelines to explain to experts how the argumentative elements were identified suggests that this is highly subjective and complex. When interacting with a simple dialogue system, students cannot debate the correctness of their answer as they could with a human partner; therefore, and to spare students the frustration of being corrected when they are technically right (Kulatska, 2019), the system's data needs to be simplified to increase the likelihood of feedback being appropriate.

The annotations would be used by the system as a gold standard to compare students' answers with. Annotating units smaller than a paragraph would require a good knowledge of syntax - this level could be expected from expert annotators, but perhaps not from high-school students who may never have taken a Linguistics class.

Using a large unit like the paragraph reduces the time needed to annotate the texts and thus enables us to produce a larger dataset, which may be more useful than a smaller one. It must also be noted that the SAT Essay texts are longer than the texts annotated in most argument mining datasets. Moreover, paragraphs are normally separated by paragraph breaks, a visually noticeable separation that could make it easier for students and annotators to distinguish between units in the text.

The author of the text divides it into the paragraphs they see fit; therefore, this division can be expected to closely match the author's idea of which elements form the text.

6.1.4 Assigning labels to paragraphs

Ideally, for the data to be simple and thus ease the annotation task, the design of the system and the ulterior educational task, each paragraph of the text should be annotated with just one label. For example, if a paragraph includes a counterargument to the author's thesis and immediately debunks it, adding a "rebuttal" label to the "argument" label would serve no purpose: the student would only be asked to spot the arguments, so that is all the annotation that the paragraph would need; other labels would only be necessary in paragraphs performing functions different from what the student would have to spot, so that the system could better explain to the student why they pointed to the wrong paragraph. However, it is not always possible to assign only one label to a paragraph. In the dataset annotation process, the following cases have been found where a paragraph may perform more than one function:

A paragraph contains both a counterargument and an argument that addresses that counterargument: it is easier to consider that counterargument as an element to back the argument, rather than another isolated argumentative function. Thus, despite the paragraph performing more than one function, we assign it only the "argument" label. As we explained in section 6.1.2, this is what happens in the December 2016 text, where, in the same paragraph, the author mentions how some people believed unpaid leave to be bad for business, and then gives statistics that prove that to be false. In some cases, however, the rebuttal and the author's argument addressing that rebuttal are written in separate paragraphs; then the paragraph with the rebuttal is given the “rebuttal” label. This is the case of paragraph seven in the May 2019 text, as we detailed in section 6.1.3 and as can also be seen in Appendix C.

A paragraph includes more than one distinct argument: when those arguments are very similar, we prefer to group them to simplify the task at all levels (annotation, development of the system, completion by the student). An example can be found in the May 2019 text (included as Appendix C), where the author talks in favor of companies disclosing gender gaps. In paragraph nine, they mention a company where this measure was implemented, and this helped them see why there was a gender gap, so they took measures that addressed this issue and improved the situation. Here we could distinguish two arguments: firstly, that disclosing gender gaps allows companies to analyze why the issue exists; secondly, it allows companies to solve the issue. However, as these two arguments are so interlinked, it is easier to consider them as only one. Still, it is possible for arguments in a paragraph to be very different, and so the paragraph will be annotated with more than one argument label. This is the case, for example, in the October 2017 text. There, the author speaks in favor of taking political action against unemployment. In paragraph four, they present their two main arguments: reducing unemployment reduces poverty and helps the overall economy of the country. These could be considered part of one single argument, as poverty levels can be considered a macroeconomic measure. However, the author seems to separate these two ideas, for example by mentioning how poverty affects children, thus making the poverty argument more linked to emotions, separating it from macroeconomic arguments.

The author presents their thesis at the beginning of the text: the thesis then needs to be accompanied by a description of the issue for the reader to understand it, and so the paragraph is annotated with the labels of those two functions. This is what happens, for example, in the October 2018 text. There, the author begins the text with a paragraph that essentially summarizes the whole text, saying that bats are dying out because of disease and accidents on wind farms (the issue), and that letting bats disappear harms agriculture and human health (the thesis).

6.1.5 Paragraph function summaries

In our annotation of the texts, aside from labelling each paragraph’s function, we decided to add a summary of the function. We can look at the December 2017 text for an example. There, the thesis (the second-hand clothes market is not as ethical as it might seem) can be found on paragraph three. Thus, before the thesis, there are two distinct paragraphs, but they perform the same function of introducing the issue being discussed in the text by saying that thrifting or donating clothes seems like a “win-win”, both cheap and ethical. As the two paragraphs are essentially transmitting the same idea, we can summarize them both as “The second-hand clothing market seems good for everyone, as it is an ethical choice and gives us cheap clothes”. After the thesis, the text contains seven other paragraphs. However, they can be summarized as three arguments: “The second-hand clothes market is more business than charity”, “The second-hand clothes market harms clothing manufacturers, especially in Africa" and "Donating clothes doesn't end the problems of fast fashion, but it gives us a false idea that it does".

We surmised that adding these summaries might be useful given the difference in length between the paragraphs and students' potential answers. Perhaps comparing the answers against data of such a different length would not give the system enough information to evaluate the answers: we decided to use a semantic similarity tool, as we justify in section 6.2, and this tool was sensitive to text length. This precaution was later shown to be unnecessary, but it did help expedite the annotation process by helping the annotator recall the content of a paragraph quickly when referring to it to create answers.

6.1.6 Assessment criteria for student answers

As we explained in section 6.1.1, we not only annotated the structure of SAT texts, but also wrote possible answers that students might give if those texts were used for ST4. We wrote those answers to analyze how a dialogue system might evaluate them to return appropriate feedback. As we mentioned, creating the answers allowed us to control the variables that we wanted to analyze - the criteria for what type of feedback would be required for each answer. The assessment criteria for which we labelled the answers were: whether the students attempt to complete the task at all, whether the students use their own words, whether the answer is complete, and whether it is correct. These are explained below.

- On-task answers

The first of these criteria should require little explanation: before the system can evaluate how the student is carrying out the task, it must detect whether they are indeed attempting to complete it. Only after the system knows that the student's answer is actually an answer can it decide whether it is good or whether the student needs some corrective feedback. For example, if the student says "I need another explanation", it would be useless to evaluate how this utterance summarizes the thesis and arguments of the reference text; the appropriate policy would be for the system to explain the task again.

- Answers in the student's own words

A good answer reflects that the student has understood the text, so that they will be able to use the information from the text for later stages in the task. A way of reflecting that one has truly understood some idea and is ready to evaluate it is to express it in one's own words (Skidmore, in Mercer et al., 2019); therefore, one important aspect of a good answer is whether the student used their own words or copied the information in the text author's voice. For example, if the text (in this example, the December 2016 text) says "The FAMILY Act is a commonsense measure whose time has come to modernize the workplace to reflect the changing face of the American family, so that finally, we can relegate having to make the choice between earning a paycheck and caring for a loved one to the dustbin of history where it belongs, a relic of a 'Mad Men' era gone by.", and the student answers “The time has come for the FAMILY Act to modernize the workplace to reflect the changing face of the American family, so that finally, we can relegate having to make the choice between earning a paycheck and caring for a loved one to the dustbin of history where it belongs, a relic of a ‘Mad Men’ era gone by. ”, the system needs to encourage them to analyze the ideas and put them into their own words.

- Complete answers

ST4 is to serve as basis for the following stage in the task, where the student will have to argue their position in an issue; argumentation is also a key aspect of learning in dialogic pedagogy, as it is about exchanging views and evaluating ideas (Andrews, 2005), two processes that are key in dialogic teaching (ibid). For these reasons, it is important that students’ answers are complete, meaning that the claims are backed by at least one premise, as giving reasons for a claim is both an indicator that dialogic teaching is taking place (Sedova, 2017) and a subskill involved in the development of oracy skills (Mercer et al., 2017). Backing claims with reasons is also a rule for discussions that students need to learn and reinforce for them to fully benefit from dialogic tasks in class (Mercer et al., 2010a). Continuing with the December 2016 text as an example, there the author defends the FAMILY Act, which would give workers paid leave for family reasons, claiming, among other arguments, that it is not fair for workers to have to choose between their family and their income, or that unpaid leave harms women disproportionately. Thus, if a student simply answers “The FAMILY Act is an important measure”, the system will need to ask them to provide some justification.

Answers that summarize other elements (aside from the thesis and arguments), to provide more context for the claim and the premises, can also be considered correct, as long as they are not merely a lengthy summary of the whole text – if the student answers giving all the information of the text, this might be a sign that they cannot distinguish the different elements of the text. Such lengthy answers have not been included; due to lack of time, other answer types have been prioritized which we believed more realistic.

Where there is no evident hierarchy among the arguments, we need to accept answers that provide only one of those arguments; demanding more of the student would require them agreeing with the annotations in the dataset used by the dialogue system on how the argumentation of the text is structured. However, argument annotation is a complex task where agreement is very difficult (Gurevych, 2021), so we need to keep the task simple to avoid correcting the student on something caused by differences in subjectivity rather than an objective error (Kulatska, 2019). Moreover, a simple system would not be able to debate with the student whether other options might be correct (such capabilities would be ideal, but they are beyond the scope of our simplified system); therefore, the task and the feedback need to be simplified. Only when the author favors one argument very visibly can the system correct the student with some confidence for choosing a different, less relevant, argument. For example, if we look again at the December 2016 text, there the author uses several arguments to defend the FAMILY Act, giving workers paid leave, without apparently giving any of the arguments special attention over the others. Thus, an answer like “The text defends a law that would allow people to get paid leave to take care of their family, so that people don't have to chose between their family and an income" could be considered perfect, even though it does not mention other arguments, such as the fact that other countries have paid leave for workers or that similar measures have not harmed businesses in the past. An example of a text where a hierarchy of arguments is clearer is the March 2019 text, where the author claims that stricter safety regulations are needed for the transportation of oil by train. The author uses several arguments, but the one that they use in most paragraphs, one of them the first one, together with the thesis, is that not requiring oil companies to remove volatile gases before transportation can lead to deadly accidents. Therefore, if the student answers "Laws should force oil companies to remove gases from the oil that is transported by train. This process wouldn't bee to difficult.", they would correctly be mentioning one of the author's arguments, that removing volatile gases is a simple process, but the system would need to help the student notice that there is a more important argument that they have not mentioned.

- Correct answers

Lastly, it is also important to remember that an argumentative text is not only a set of structural elements, but that these elements reflect a stance. For instance, in the annotation guidelines by Stab and Gurevych (2014), identifying the author's stance is seen as a crucial task that is to be completed before analyzing the structure of the text, and it informs the identification of the different structural elements. The students' understanding of the author's stance is thus considered an important part of the educational task - their answer has to not only deal with the same issues as the author's thesis in the text, but also discuss them in the same positive or negative light. Reflecting a different stance could be a sign that the student has misunderstood the text, so they might not be able to evaluate its ideas and bring them into discussion. For instance, we can look again at the March 2019 text, where the author says that oil companies need to be forced to remove volatile gases before transportation in order to avoid deadly accidents. There, if a student answered "The new safety regulations for transporting oil don't require companies to remove volatile gases. This was an acceptable decision - gases are not the main problem.", they would have completely misunderstood the text and would need some guidance.

6.1.7 Answer distribution

So that the dataset can serve to train the system to classify answers as good or bad according to our established assessment criteria, it would need to contain some perfect answers, as well as answers lacking in some of those criteria. A complete dataset could feature all possible combinations of criteria (i.e. perfect answers, answers lacking in one aspect, answers lacking in two, answers lacking in three, answers lacking in all four aspects). However, for a scaffolded task, feedback can focus on one aspect, so that the student can work on one skill at a time. For this reason, it is necessary to establish a hierarchy in the assessment criteria, so that the system can focus on the most important aspect when problems are detected concerning more than one assessment criteria. Naturally, the first step is determining whether the student's utterance is an answer. After that, we have established a hierarchy with completeness as the first aspect to be assessed, followed by paraphrasing, and ending with correctness:

1. Completeness
   The task preceding the summarization of the thesis and the premises involves spotting the paragraphs where that information is conveyed. Thus, completeness could be the first aspect to be assessed: if the student does not include all the elements they spotted in the previous stage, that stage may need to be reviewed and the student reminded of the most important text chunks.

2. Paraphrasing
   Once the student knows what information they need to convey, we would need to assess whether they understood the ideas contained in the paragraphs they pointed to: this would involve both paraphrasing and accurately conveying the author's stance. As putting other's ideas into one's voice can be seen as a vehicle towards incorporating those ideas into our understanding (Skidmore, in Mercer et al., 2019), we could consider the paraphrasing aspect earlier in the hierarchy. This would leave the aspect of correctness third and last in the hierarchy.

3. Correctness
   A student who has understood which bits of information they need to use and who has put that information into their own words could be expected to only rarely make any errors in interpreting the stance conveyed in what they just read. Still, this possibility has to be contemplated. Such cases might reveal that the student's reading comprehension is not at the level of the task, and perhaps it might be useful to add a multiple-choice subtask that brought the student's attention to specific points in the text and helped them see which stance they convey. Opting for the multiple-choice format with no Natural Language Understanding could make the integration of the subtask easier (Jokinen, 2009), though it would require designing the questions and possible answers for each text used by the dialogue system. Perhaps a good aspect-based-sentiment analysis tool could help automate this task, as the tool could recognize which expressions conveyed a positive or negative stance and bring the student's attention to those specific parts of the text. That, however, is to be analyzed in future studies.

With this hierarchical approach in mind, as well as to also ease our dataset compilation efforts, we have only fabricated perfect answers and answers lacking in one aspect (i.e. answers that are off-task, answers that are incomplete, answers that are incorrect, and answers that are copying the text instead of paraphrasing it). We acknowledge the limitations of this approach: though the system is meant to focus on one assessment aspect at a time, real student answers may present issues with more than one assessment criteria. Detecting the issue highest in the hierarchy in such answers that present more than one problem might complicate the classification task for the system: e.g. if an answer is a poor paraphrase will the system also detect that it is incomplete and apply the policy for returning feedback concerning completeness, the higher aspect in the hierarchy? Nonetheless, we believe it sensible to start our tests with simpler data and only test how more complex answers are classified once we confirm whether simple answers can be accurately classified.

We also need to establish the criteria for the design of the answers, beyond the type of error they may feature, as that is too vague a guideline - the dataset could become very imbalanced and it might be hard to extract clear conclusions from an experiment. Below we list the labels for each answer type and explain how answers with that label were created:

- Off-task questions (offt): these are questions that do not attempt to complete the task, either because the student is confused or because they do not want to do the task.
  - offt-needhelp-questask: Direct question about task instructions (e.g. "What do I need to do now?").
  - offt- needhelp-commtask: Command to clarify something (e.g. "Gimme some help").
  - offt- needhelp-conf: Expression of confusion (e.g. "Eh?").
  - offt-gibb: Random utterance, here in the form of random song lyrics (e.g. "Heart beats fast Colors and promises How to be brave? How can I love when I'm afraid to fall?").
  - offt-rand: Random/off-task utterance, but with at least one word from the domain of the reference text (e.g in the December text about the FAMILY Act, "Never heard of this FAMILY Act").
  - offt-disg: Expression of disgust at task (e.g. "This suuuckss").
  - offt-pers: Question about the system's personality (e.g. "Are you an evil robot?"). Bii et al. (2013) observed that these occurred with some frequency among high-school students using a chatbot.
- Perfect answers (pa): answers that are complete (state the author's thesis and at least one way it is justified), good paraphrases (they do not copy the text, nor do they simply replace or move a few words), and that are correct (state the author's position on the issue, and not an opposed view). For example, for the December 2017 text about how the second-hand clothes industry is not as ethical as it seems, "Most people believe that donating clothes is a good thing, but they're harming some businesses and the planet cos it's just putting a patch on fast fashion, not solving anything" would be a good answer.
- Poor paraphrases (poorpar): answers that take a text chunk containing the thesis and simply move some words or replace them with synonyms to make it less of an exact copy. When the thesis is stated in more than one paragraph, we take them all as reference for an equal number of answers (e.g. if the thesis is in two paragraphs and a dataset category is meant to have four answers, half will poorly paraphrase one

paragraph, and the other two answers will poorly paraphrase the other paragraph). The poor paraphrases that used synonyms were created using Wordnet to expedite the process, despite this resulting in strange utterances (still, replacing words with inadequate synonyms is something that students might in fact do, if perhaps not to the same extent).

- poorpar-syn2: Replacing 20% of words in the paragraph with synonyms. A short example that we can show here comes from the May 2017 text, where the thesis can be found on paragraphs eight and twelve. Paragraph twelve says "With so much to gain, we need to cut work hours while there is still time.", and with 19% of the words replaced it becomes "With therefore much to profit, we motivation to cut work hours while there is still time".
- poorpar-syn5: Replacing 50% of words in the paragraph with synonyms. The previous example paragraph would now become "With then much to addition, we necessitate to hack work hour while there be still time" - it was not always possible to find synonyms for half the words, especially in short paragraphs.
- poorpar-ord1: Making a small change in the paragraph's structure. Continuing with the May 2017 example, paragraph twelve becomes "With so much to gain, we need to cut work hours", where we merely removed the end of the sentence.
- poorpar-ordmore: Making several changes in the paragraph's structure. For example, here paragraph twelve becomes "We need to cut work hours. There is much to gain, but we need to make sure there is still time.", where we reordered the sentence and changed the connectors between the sentence parts.

• Incomplete answers (com): Answers that are missing the text thesis or its justification, or both (but are genuine attempts at completing the task). To illustrate what types of answers are assigned this label, we will be using examples from the October 2017 text, where the author defends taking political action against unemployment to reduce poverty and improve the economy.

  - com-nojus: Stating the author's stance without justification (e.g. "Congress needs to pass a bill that reduces unemployment").
  - com-offfocus: Summarizing text parts that merely provide context (e.g). "FDR signed the so-called second bill of rights to reduce unemployment".
  - com-minarg: Stating the author's thesis but justifying it with a minor argument. This category only applies to texts with a clear hierarchy of arguments. An example would be "Policy makers need to tackle unemployment. That could also solve infrastructure issues, killing two birds with one stone".
  - com-arg: Stating an argument without linking it to the thesis (e.g. "If people have no job they can't feed their family").

  - com-topic: Describing the topic without stating a position (e.g. "The text discusses the consequences of unemployment and what Congress can do about it").

- Incorrect answers (corr): Answers that contain a thesis and justify it, but these elements contradict the author's position. For this category we'll be showing examples from the October 2018 text, where the author claims that we should protect bats from disease and accidents, because bats provide excellent pest control for agriculture and disease control.

  - corr-negthes: Inverting the author's stance (e.g.) "We shouln't put our efforts into protecting bats, because we have other means of killing insects.".
  - corr-negarg: Stating the author's thesis but negating the justification (e.g. "Pesticides are safer and more effective than bats for getting rid of pests in agriculture, but we still gotta preserve them as any other endangered species.").
  - corr-negall: Inverting the author's thesis and its justification (e.g. "It appears that bats are dying out, but humans are in no position to solve this. Also, bats are a very harmful species, so it's no big loss.").

The distribution of answer types is shown in Figure 2. The first decision was which proportion of answers would need to be on- and off-task. Preliminary analyses suggested that distinguishing between on- and off-task answers could be easy, and thus fewer off-task answers might be needed. Inside the category of off-task answers, all sub-categories were given a similar weight, giving only slightly more weight to the off-rand category (where off-task answers contain a domain word), to have sufficient data to draw conclusions on the system's reliance on single appearances of domain words. On-task answers include both perfect answers and answers with one problematic aspect. Inside each of the three categories of aspects that we considered for feedback there were many subcategories; therefore, to have enough data to draw conclusions on each of those sub-categories, each of the three categories had to be assigned a sizeable proportion of the dataset. This left the category of perfect answers with only 10% of the total; however, as it has no subcategories, this might be sufficient. The completeness category is the one of the three feedback categories with the most subcategories, accounting for the main ways in which a student may fail to include a thesis and a premise to back it in their answer. For this reason, as well as the fact that completeness has been assigned first place in the hierarchy of feedback types, this category has been assigned the largest proportion of the dataset. The proportions assigned to the other two feedback type categories respond to the same reasoning. Firstly, whereas completeness has five subcategories, paraphrasing has four, and correctness three. Also, in the hierarchy of feedback types, completeness was put in first place, paraphrasing in second, and correctness in third.

We must acknowledge that this category distribution is not perfectly balanced; some categories have been assigned a larger proportion of the answers when we believed that classifying those categories would be more challenging. To be able to create a number of answers that could make the dataset usable (i.e. allowing us to extract some initial conclusions, despite how limited they might be), we needed to simplify the creation process. Real student answers could be expected to belong to the negative class for more than one of our classification criteria (on-/off-task, completeness, correctness, paraphrasing). However, creating more realistic answers that failed in more than one aspect while maintaining a balanced distribution would have slowed down the fabrication of answers and reduced output significantly (e.g. creating incomplete answers that only contained the thesis and no argument was easy, but having to make some of them be also bad paraphrases while keeping count how many answers need to be in each class would have required much more time or more annotators). Nonetheless, as will be discussed in section 6.2.3, the category distribution does not seem to have a strong impact on our tests.

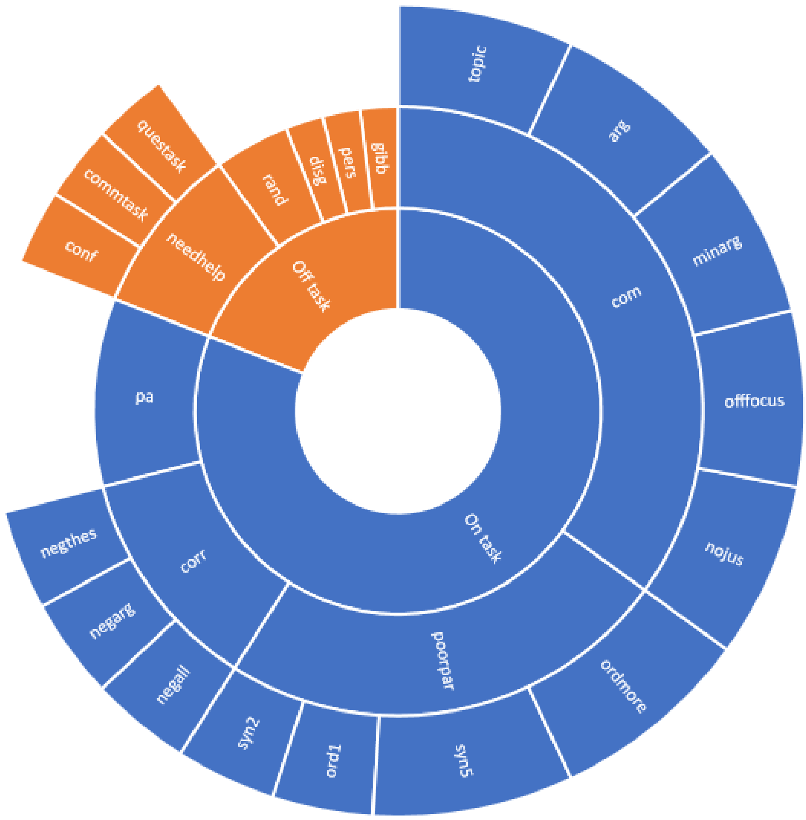


Figure 2: Category distribution

6.1.8 Dataset size

The size of the dataset was decided merely based on what was feasible in the time available. In the end, we were able to annotate ten SAT Essay texts for their argumentative structure. For each text, 93 or 100 answers were created - one answer type was not applicable to all texts, as explained in section 6.1.7, resulting in six texts having 93 answers instead of 100.

We acknowledge the limited size of the dataset, but, as will be discussed in sections 6.2.3 through 6.2.5, this was enough to reach some initial conclusions from an experiment.

6.2 Experimental tests

Here we describe the classification experiment that we carried out to determine the technical requirements to start developing ST4. Our goal is to determine whether a tool that we have selected (detailed below) and other information that we extract from our dataset could be enough for a dialogue system to classify student answers according to our assessment criteria (section 6.1.6). Below we provide an overview of what our tests consisted on. Then, in section 6.2.1, we describe in detail the features that were used in the classification tasks. Then, section 6.2.2 describes those classification tasks. After that, in section 6.2.3, we discuss the main results of our tests. This is followed by a more detailed analysis, in section 6.2.4, and a discussion of what these results would imply for the design of the system, in section 6.2.5.

As ST4 is concerned with analyzing students' paraphrase of the main ideas of a text (thesis + arguments), we hypothesized that comparing the answers against the relevant paragraphs could provide the system with the necessary features to evaluate the answers. The tool used for the comparison was a Semantic Textual Similarity (STS) tool developed by the IXA group using a model by Cer et al., the Universal Sentence Encoder. The model converts sentences into vectors, so the similarity between sentences can be calculated by comparing their vectors (ibid). Cer et al. demonstrated that their Universal Sentence Encoder can be used for several NLP tasks with good results. In our experiment, we used the STS tool to compare each answer against each paragraph of the text for the exercise that the answer responded to. Answers were also compared against the paragraph function summaries. This way, we were able to obtain scores that reflected how similar a student answer was to each paragraph and their summaries. These scores could then be used to classify the answers according to our assessment criteria (being on task, being a good paraphrase, being complete and being correct, as described in section 6.1.6). To do this, we performed four classification tasks, each focusing on labelling answers according to one of the criteria. For each task, we trained a classifier using the STS scores data, together with some additional information which we discuss below, for part of the answers and test on the remaining answers whether the model learned by the classifier performs adequately.

6.2.1 Features used for classification

The dataset contains answers for ten different texts. These texts have very different structures. In order to test a classifier that could work for any of the texts (texts from our dataset and future texts), we needed the classifier models to use features that were applicable to any of the texts, regardless of structural differences - for example, if the classifiers learn to classify answers for texts where a counterargument is presented before the main argument but not for texts where no counterarguments are mentioned, they will be of little use. For that reason, we created some compound features, turning the varying number of

STS values into a constant number of values. The resulting features for each answer of each text were as follows:

1. Average STS value of comparing an answer with all the paragraphs containing the thesis
2. STS value of comparing an answer with the thesis summary
3. Average STS value of comparing an answer with all the paragraphs containing arguments
4. Average STS value of comparing an answer with all argument summaries
5. Average STS value of comparing an answer with all paragraphs
6. Maximum STS value of comparing an answer with a paragraph containing the thesis
7. Maximum STS value of comparing an answer with a paragraph containing an argument
8. Maximum STS value of comparing an answer with an argument summary
9. Maximum STS value of comparing an answer with any paragraph
10. Answer word count
11. Unigram overlap between the answer and the whole text
12. Bigram overlap between the answer and the whole text

These last three features, as can be seen, are not related to STS scores, but were added because they seemed potentially very useful and required no more than a simple script to be obtained. We believed that the overlap measures would be especially useful for the classifier evaluating answers with regard to whether the answer was in the student's own words or whether it was a poor paraphrase - if the answer has many words in common with the text, then these are not the student's words. We also measured word count thinking that it might be a useful feature for all classifiers, but especially for the one evaluating whether answers are complete - extremely short answers might be less likely to contain the thesis and an argument.

It seems necessary to explain how unigram and bigram overlap measures were calculated, so that the implications of our chosen method can be taken into account. We removed punctuation, as overlap in this respect between the reference texts and the answers would not provide any information useful to our task. To calculate unigram overlap, we also removed stopwords, for the same reason. However, to calculate bigram overlap, we did not remove stopwords, for fear of that resulting in artificial bigrams - technically, removing punctuation could also result in artificial bigrams, but we assumed that the noise removed would be greater than the noise introduced. For example, we feared cases where an answer said something as "volatile gases are bad", and the reference text said "The situation is volatile, as the gases in oil..."; removing stopwords could introduce the false bigram "volatile gases". Future tests with bigram overlap, and tests on larger n-grams, may reveal whether our precautions were excessive and introduced much useless noise, or whether most noise was as bigrams formed by terms already counted in the unigram overlap together with a function word (e.g. "the volatility", instead of a "true" bigram like "gas volatility"). Whichever the case, noise that simply increases bigram count unnecessarily but does not introduce artificial bigrams may result in no more issue than needing to use a higher threshold to classify answers based on the bigram overlap feature.

6.2.2 Classification tasks

Once all the features were obtained, as explained in 6.2.1, we performed a series of binary classification tasks. The first classification was to divide answers as on- or off-task. Secondly, we tested how answers could be classified as complete or incomplete (off-task answers were also assigned the "incomplete" target label, together with the answers in the specific "incomplete" category described in section 6.1.6). Another classification test was performed to detect whether answers were in the students' own words or whether they were loosely copying the text (only answers in the specific "poor paraphrase" category described in 6.1.6 were assigned the negative target label). Finally, answers were to be classified as correct or incorrect (the negative target label was assigned to both the answers in the specific "incorrect" category described in 6.1.6, as well as the "off-task" answers).

The classification tests were performed using the scikit-learn decision tree classifier tool due to the interpretability of results it offers. The classifier parameters were selected with the goal of preventing overfitting by controlling the size of the tree. Tests were performed with ten-fold crossvalidation using 20% test size. We also performed tests dividing the data by document with a simple 80%/20% train/test split, without k-fold crossvalidation, as this allowed us to see individual decision trees, instead of only the general metrics obtained with the crossvalidation tests. The split was random, but coincidentally the two texts of the test set were one with an argument hierarchy and one without, resulting in a balanced sample.

6.2.3 General findings

Table 1 shows the accuracy scores obtained by the decision tree classifier for each of the four binary classification tasks (the scores from the second column are averages for the ten crossvalidation folds, together with the standard deviation (SD)). The tasks are indicated on the left column. Each of the other columns corresponds to a method used to split the dataset into a training a testing subset. As explained in section 6.2, we used ten-fold crossvalidation, as well as a simple split of 80% of documents for training, 20% for testing, to see more clearly the impact of individual texts' characteristics on results. The resulting split was balanced with regard to the number of documents with a hierarchy of arguments (as explained in 6.1.2, some texts present arguments of equal weight, while others have a

CLASSIFICATION TASK	10-k CV (Avg.)	Split by document
ON/OFF TASK	0.96 (SD:0.01)	0.96
COMPLETENESS	0.85 (SD:0.02)	0.77
OWN WORDS	0.98 (SD:0.01)	0.99
CORRECTNESS	0.85 (SD:0.02)	0.83

Table 1: Accuracy scores for each classification task with two dataset splitting methods

hierarchy).

What we can first observe from these accuracy scores is that the features we have used in this initial test might suffice for some of the classification tasks. The easiest classification tasks seem to be determining whether answers are on or off-task and whether they are good paraphrases: nonetheless, tests with real answers would probably return lower accuracy scores, at least until enough real answers can be gathered to train a new, more sophisticated classifier. We have striven to write diverse, plausible answers, but we cannot anticipate what real answers may truly be like. Students' bad paraphrases may perhaps be less bad than the ones we created with Wordnet, which would complicate the classification. The scores for the other two classification tasks, the ones concerning completeness and correctness, while good, are not as high, even with simple, artificial data. This suggests that the features used for classifying answers with regard to these categories may not be enough; other more adequate models may be required to extract more useful information from students' answers for their classification (e.g. models trained for stance detection might help determine whether students' answers reflect the correct stance from the reference text). Table 1 also shows that classifying answers by completeness is affected by the idiosyncrasy of the reference texts, as gleaned from the lower accuracy scores obtained when the dataset was split by document.

Figure 3 shows the confusion matrices for the classification tasks (with the simple dataset split, using 20% of documents as test set). The matrices show that classification was mostly done well, with some minor exceptions. Firstly, we can observe that, overall, there were very few false negatives - very few good answers were classified as bad. The highest rate of false positives was found in the classification task according to completeness: 5% of answers from the test set. These good results could perhaps be due to the imbalanced categories. In section 6.1.7 we justified the need to work with these imbalanced categories. The scale of this project and the goal of merely getting a broad idea of the STS tool's suitability for ST4 of the task do not warrant an expansion of our experiment to try different methods of artificially balancing the sample distribution. Still, we performed one such test to see to what extent the high accuracy in some categories might be due to the imbalance in sample distribution. We took the best-performing category, paraphrasing, where results might be "too good to be true", and up-sampled the minority class (bad paraphrase). Accuracy decreased, but only by 3%, which maintains it above 90%. Thus, it seems that the high accuracy cannot be attributed solely to the imbalanced distribution, but mainly to the quality of the STS model. Nonetheless, once sufficient real answers could be gathered, more realistic class distributions could be achieved. They might still be imbalanced if students' performance is not balanced, and some types might not even appear (e.g. in a study setting, participants might be the people with the most motivation to carry out the task and everyone's performance would be excellent (Pinkwart et al., 2008)). Thus, our limited dataset at least has the advantage of containing a wide variety of answer types.

If we look at the false positives (bad answers that were classified as good), results are only slightly worse for the on-/off-task and the own words/bad paraphrase classification tasks; for the completeness and correctness classification tasks, results are noticeably worse. Almost 18% of answers from the test set were incorrectly classified as incomplete, and 15% were wrongly classified as correct. While these are not excellent results, especially considering that we used simple, artificial answer data, we believe that, in an educational task, false positives are less problematic than false negatives. If a student receives corrective feedback on their good answers, they could get frustrated and unmotivated (Kulatska, 2019). On the other hand, if a student gives a bad answer that is not so clearly bad for the system to label it as such, not giving them feedback might not be so problematic - if their answer was not extremely bad, their need for feedback may also not be so extreme.

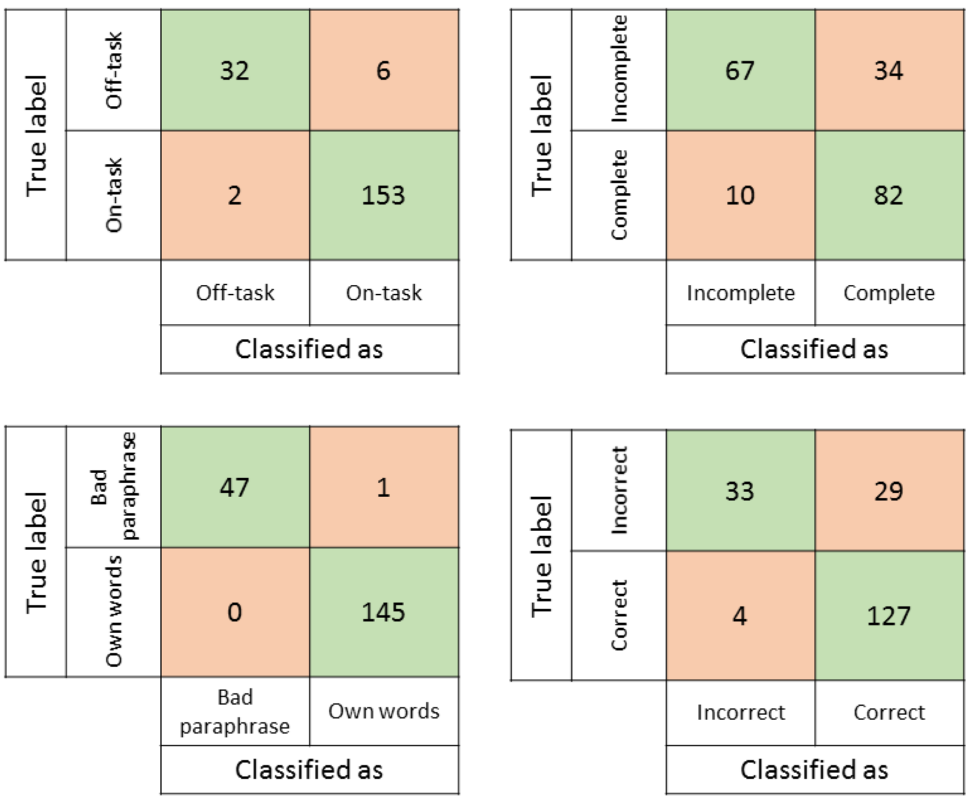


Figure 3: Confusion matrices with dataset split by document

We also analyzed the features that the classifiers found most informative for each of the classification tasks. What first catches our attention is that the function summaries described in section 6.1.5 turned out to be unnecessary for classification; at least they served to expedite the annotation process. The features most used by the classifiers were 1) the maximum STS score of an answer compared with a paragraph containing the thesis and 2) the word count. The maximum STS score of comparing answers with the whole reference text was also used to classify answers for correctness and whether they were on task. The correctness classifier also used the maximum STS score of answers compared with a paragraph containing an argument. The completeness classifiers used either the maximum of the average STS score of answers compared with a paragraph containing an argument. Bigram overlap was also used in several classifiers, most consistently to classify answers as on or off task. This classification task also used unigram overlap. This contradicts our hypothesis that overlap measures might be most useful for classifying answers as good or bad paraphrases - for that classification task, the most informative feature was the maximum STS score of comparing an answer with paragraphs containing the thesis.

6.2.4 Analysis by assessment criteria

Here we take a closer look at the results for each category. We start with the categories that performed the worst; we believe that they do not warrant an in-depth analysis, as the low performance may be due to the STS model not being ideally suited for them. Nonetheless, we take a general look at these categories and comment on how results could be improved in future tests. For the best performing categories, we do provide some more details on how the features were used for classification and provide a graphical example of how it works.

Our different splitting methods revealed that completeness is the classification task most sensitive to the reference texts' idiosyncrasy, as seen most evidently on Table 1. Further tests would need to be carried out to arrive at a better way of classifying students' answers with regard to this completeness. We surmise that a model trained to distinguish arguments from unreasoned statements might help - complete answers state a position and back it with at least one argument. However, seeing how sensitive this category is to each document's characteristics, it might still be necessary to combine that model with a similarity model like the one that we have used, or some other, to make sure that the answers are justified with the required elements from the text, and not something else (e.g. a minor argument, or some other argument invented by the student before they are asked to think of their own arguments).

Correctness, while not as problematic as completeness, was also a criteria where classification showed much room for improvement. The confusion matrices show that there could be many false positives in this category - many incorrect answers could be labeled as correct and receive no corrective feedback. We mentioned in section 6.2.3 that it might not be very problematic if some slightly incomplete answers do not receive feedback; however, me must also assume that accuracy would drop when using real student data, and false positives might reach a proportion not appropriate for a pedagogical task like the one we propose. If the students do not receive the feedback they need, they might not be ready to complete the following part of the task - we would be removing the scaffold before the student can work without it. Given that detecting whether an answer is correct or incorrect is, in essence, detecting whether the students' answers have the same stance as the author's text, a model trained for stance detection could improve accuracy.

We now move on to the paraphrasing criteria; the features that we have used seem to be adequate for obtaining good classification performance in this and the on-/off-task criteria. We include Figure 4 as an example of how classification could be done with regard to paraphrasing; we have selected this tree as the most illustrative example of how the classifiers work, because it is the simplest tree. We see that there are two very clear thresholds of similarity with the text's thesis above which the answer can be confidently considered a near copy of the text, and below which the answer can be confidently considered to be in the student's own words (off-task answers would be the ones lowest below the threshold). Nonetheless, there is a narrow band between those thresholds where classification cannot be performed so confidently - some good answers would be at this point of being very similar to the text, but not so similar that they copy it. Fortunately, this uncertainty was happily resolved in our tests and there were no false negatives (good paraphrases labelled as bad). There were, however, a small number of false positives (bad paraphrases labelled as good). An analysis of the classification criteria suggests that false positives would mainly be bad paraphrases in the form of short sentences (short instances of the poorpar-ordmore category). The STS model that we have used thus seems exceedingly sensitive to text length. A look at the longer answers confirms this assumption: very long bad paraphrases were given a similarity score of 1 (the maximum) despite not being exact copies of the text, which is what that maximum similarity score would imply. Nonetheless, the accuracy scores obtained in these initial tests suggest that this or other similar models (perhaps a model trained with sentences and paragraphs of varying length) could give adequate results.

We finally look at the on-/off-task classification, which we assumed would be the most accurate, though the paraphrasing category surpassed our expectations. We have observed a clear threshold of similarity with any paragraph of the reference text above which answers can be confidently classified as on-task; there is also a threshold for unigram overlap that results in zero entropy. However, answers below these thresholds cannot be classified so confidently based on the available features. Unigram overlap can be an informative feature for most cases, but two off-task subcategories present some problems: off-rand (off-task answers containing a domain word or expression) and off-gibb (random answers, which we created by randomly copying song lyrics from different genres). The issues with the off-rand category were expected: these off-task answers were designed to be problematic by including domain words. Therefore, they may have a higher unigram overlap than some on-task answers that use synonyms instead the specific domain words from the text. The issues with the off-gibb category were more unexpected; it seems that removing stopwords is not enough, and perhaps the overlap value should be extracted counting only domain words, which would have to be distinguished from non-domain-specific yet also

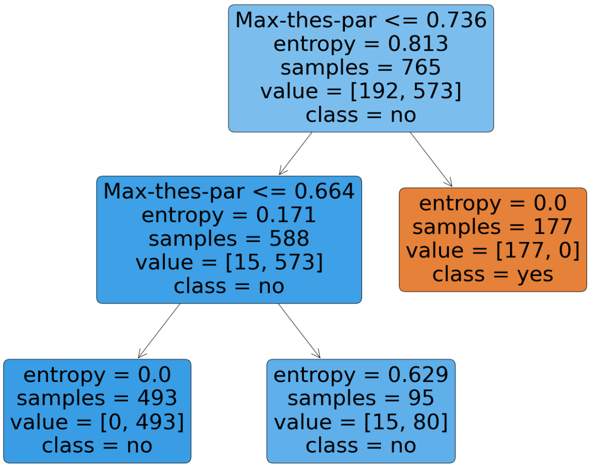


Figure 4: Decision tree example for classification into bad paraphrase/own words. Majority class is own words.

non-stopword terms. Thus, labelling the text domain (manually or automatically) might be a necessary first step, to then compare only unigrams belonging to that domain (perhaps with a domain-annotated lexical database, such as Wordnet Domains[23]). Difficulties in classifying these and other off-task subcategories also seem to stem from the STS tool's sensitivity to text length. For example, an answer like "Eh?" is given a similarity as high as 0.538 to a text paragraph, the same as an answer like "The used-clothes industry is not ethical", which is an on-task answer that is simply too short because it is missing the justification for the thesis. Nonetheless, there were only a few false positives in our tests.

6.2.5 Implications for system design

Firstly, we can conclude that annotating the paragraph function summaries described in 6.1.5 would be unnecessary, as the features obtained from them seemed to provide no useful information for classification. Still, they are useful in the annotation process to provide a quick overview of the reference text.

The classification accuracy scores obtained in the different categories suggests that the model that we have used could be adequate for determining whether students' answers are on task and whether they are good paraphrases, though it might be necessary to also procure and test a model more adapted to varying text lengths. Perhaps once a first version of the dialogue system was fully developed and sufficient real answers could be gathered,

[23]Link to Wordnet Domains

a model could be trained to obtain even better accuracy. The accuracy scores from our tests also suggest that a different model would be needed to provide sufficiently accurate classification of students' answers with regard to completeness and correctness.

If the dialogue system were to be developed without additional, more adequate models, the hierarchy that we established for the different assessment categories might have to be modified to be more in line with the system's capabilities (the hierarchy was described in section 6.1.6). Naturally, our first level in the hierarchy was determining whether the student's answer was an actual attempt at an answer; fortunately, this was done very accurately in our tests, so this level could be maintained in first place. We then set completeness as the second level of the hierarchy: as the previous subtask (ST3) required the student spotting paragraphs that contained the author's thesis and at least an argument, the next logical step might be ensuring that the student used all the elements they had identified. However, given the good-yet-not-ideal accuracy scores obtained for completeness classification, this second level might need to be switched with the third criteria: whether the answer is a good paraphrase. Paraphrasing ideas can be seen as a way of processing and incorporating them (Skidmore, in Mercer et al., 2019); checking whether this processing is taking place could be done before checking whether the student processed all the text elements that they needed to process. Correctness is the aspect where we would expect the fewest issues: a student that has produced a complete answer and paraphrased the text appropriately could be assumed to have done enough processing to not misinterpret the text. As such, correctness could remain last in the hierarchy. In addition to changing the order in which the different assessment criteria would be checked, the system's potential inaccuracy at some categories could be compensated by taking confidence into account (Jurafsky and Martin, 2019). When an answer did not meet the criteria for confident classification into a category, classifying the answer would run the risk of frustrating the student by giving mistaken corrective feedback (Kulatska, 2019) or of not providing necessary feedback to a student whose answer could need some improvement. An alternative would be to respond to such answers that cannot be classified confidently with no corrective feedback, but some other non-committal prompt that could guide the students who needed guidance, but which could be ignored by students who gave a perfect answer (e.g. "Not bad :) Did you check if your answer contains the author's thesis and a supporting argument from the text? I forgot to check"). To ensure that the student is given enough scaffolding, it would also be possible to add a simple multiple-choice task that checked students' grasp of the aspects that the system could not evaluate confidently - though this would imply creating such questions, either manually or automatically.

7 Conclusion

In this project, we have extracted valuable insights from the literature on dialogic teaching and the use of dialogue systems and other technologies in education, and have translated them into several contributions which we hope can pave the way for pedagogically-motivated systems that can assist in the implementation of dialogic teaching. As we mentioned in sections 2.2.1 and 2.2.2, dialogic teaching is a pedagogic approach that shows great promise for improving students' academic performance (Jay et al., 2017), and which reinforces key skills needed in everyday life (Mercer et al., 2017); however, implementing this approach can be challenging (Mercer et al., 2010b; Sedova, 2017), but technology may be of some help (Major et al., 2018).

Our first contribution has been linking findings from the literature to dialogue system features (section 4). This link would ensure that the ensuing system is designed specifically for dialogic teaching, and it could thus be more adequate than other tools which are used but were not designed for this or any other pedagogic approach (Mercer et al., 2010b). Nonetheless, as we mentioned in section 1, our hypothesis that dialogue systems can be useful for dialogic teaching cannot be fully confirmed until such systems are tested with a student sample. To pave the way for these dialogue systems to be developed and tested, we have made a second contribution as a proposal for an argumentative task to be completed using a dialogue system (section 5). We have detailed the features of the system that would carry out the proposed task (sections 5.2 through 5.7). In addition to describing the system's proposed features, we have shown more specifically which form they could take, through a detailed list of dialogue acts (section 5.6 and Appendix A) and a possible example conversation (section 5.7 and Appendix B). We have also explained how the proposed features would comply with our proposed framework 5.8. Our third contribution takes the form of a dataset 6.1 and the tests that we performed with it (section 6.2) to see how one of the core subtasks from our proposed task could be implemented.

Our initial experiment has shown that a semantic similarity model could suffice for part of the subtask, especially if some adaptations are made to compensate for technical limitations, at least until better data can be obtained to train and test better models. It seems that determining whether students' utterances are attempts at an answer to the proposed task, as well as whether they are acceptable paraphrases of the source text and not near copies, could be relatively easy. More tests with different tools on this and other data would be needed to decide the best way to assess students' answers with regard to other assessment criteria, as well as to analyze the system requirements for the remaining parts of the proposed task.

7.1 Limitations and further work

The fact that we were only able to focus on ST3 and ST4 constitutes a limitation. However, as we justified in section 6, we believed this to be the subtasks for which we could contribute the most. Moreover, as our proposed task is designed to provide scaffolding to the student, a system that reached only ST4 might still help the students develop their oracy skills and prepare them for carrying out ST5 with a human partner instead of a dialogue system.

The characteristics of our dataset (6.1) also present some limitations. One of them is the distribution of answer types. As we explained in section 6.1.7, we planned the distribution aiming to simplify the creation of the dataset and to make the results of our initial tests as informative as possible. This resulted in imbalanced class distribution for our experimental classification tasks. Nonetheless, as mentioned in section 6.2.3, this did not seem to have a very strong impact on our results. Our dataset is also limited with regard to its size and the fact that it consists of artificial answers. However, our data has proven useful for basic tests, which can inform future studies that gather more data of better quality. Despite the answers being artificial, the annotated texts from the dataset are real argumentative texts selected and adapted by experts for their use in tasks aimed at high-school students (CollegeBoard, 2015); thus, this part of our dataset at least could be considered of more quality and value, and might be used to carry out a Wizard-of-Oz study to gather data in the form of student answers.

Despite these limitations, we hope that these can indeed be the first steps towards a successful dialogue system. We have provided design details and examples for the development of a pedagogically-motivated argumentative dialogue system that could help in the implementation of dialogic teaching. In addition to our proposed task, we have also contributed a more general framework that we hope can inform the design of any other dialogue system intended for dialogic teaching.

References

S. Abbasi and H. Kazi. Measuring effectiveness of learning chatbot systems on student's learning outcome and memory retention. *Asian Journal of Applied Science and Engineering*, 3(2):251–260, 2014.

R. Alexander. *Dialogic teaching essentials.* National Institute of Education, Singapore, 2010.

R. Andrews. Models of argumentation in educational discourse. *Text*, 1(25):107–127, 2005.

P. Bii, J. Too, and R. Langat. An investigation of student's attitude towards use of chatbot technology in instruction: The case of knowie in a selected high school. *Education Research*, 4(10):710–716, 2013.

F. Catania, M. Spitale, G. Cosentino, and F. Garzotto. Conversational agents to promote children's verbal communication skills. In *Proceedings of CONVERSATIONS 2020 – the 4th International Workshop on Chatbot Research,*, pages 158–172, Amsterdam, 2020. Springer LNCS.

D. Cer, Y. Yanga, S. Konga, N. Huaa, N. Limtiacob, R. St. Johna, N. Constanta, M. Guajardo-Cespedes, S. Yuanc, C. Tara, Y. Sunga, B. Stropea, and R. Kurzweila. Universal sentence encoder.

H. Chen, X. Liu, D. Yin, and J. Tang. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35, 2018.

CollegeBoard. Suite of assessments annual report, 2020.

CollegeBoard. Test specifications for the redesigned sat, 2015.

H. Douglas Brown. *Language Assessment: Principles and Classroom Practices.* Pearson, 2004.

Y. Goda, M. Yamada, H. Matsukawa, K. Hata, and S. Yasunami. Conversation with a chatbot before an online efl group discussion and the effects on critical thinking. *The journal of information and systems in education*, 13(1):1–7, 2014.

Google. Conversation design, 2021. URL `https://developers.google.com/assistant/conversation-design`.

I. Gurevych. Let's argue: Understanding and generating natural language arguments, 2021. URL `https://www.youtube.com/watch?v=zuz1cIyB-XU`.

B. Heller. Conversational agents as historical figures: Individual differences and perceptions of agent and social presence. *EdMedia+ Innovate Learning*, pages 1374–1380, 2016.

W. Huang, D. E. Gonda, and Khe F. Hew. Designing and evaluating three chatbot-enhanced activities for a flipped graduate course. *International Journal of Mechanical Engineering and Robotics Research*, 8(5):813–818, 2019.

IBM. Project debater, 2021. URL `https://www.research.ibm.com/artificial-intelligence/project-debater/`.

T. W. Jay, B. Willis, P. Thomas, R. Taylor, N. Moore, C. Burnett, G. Merchant, and A. Stevens. *Dialogic Teaching: Evaluation Report and Executive Summary.* Education Endowment Foundation, London, 2017.

K. Jokinen. *Dialogue Systems. En K.* Speech Interaction and Rational Agents (págs. 99-112). Wiley, Jokinen, Constructive Dialogue Modelling, 2009.

D. Jurafsky and J. Martin. *Dialogue Systems and Chatbots.* (In press), Stanford, 2019.

I. Kulatska. Arguebot: Enabling debates through a hybrid retrieval-generation-based chatbot. Master's thesis, University of Twente, 2019.

N. L. Kuyven, C. A. Antunes, Barros Vanzin V. J., Tavares da Silva J. L., Loureiro Krassmann A., and Rockenbach Tarouco L. M. Chatbots na educação: uma revisão sistemática da literatura. *RENOTE*, 16(1):123–132, 2018.

L. Major, P. Warwick, I. Rasmussen, S. Ludvigsen, and V. Cook. Classroom dialogue and digital technologies: A scoping review. *Education and Information Technologies*, 23(5): 1995–2028, 2018.

M. Marge, C. Espy-Wilson, and N. G. Ward. Spoken language interaction with robots - research issues and recommendations. Dallas, 2020. NSF Future Directions Workshop.

N. Mercer and C. Howe. Explaining the dialogic processes of teaching and learning: The value and potential of sociocultural theory. *Learning, culture and social interaction*, pages 12–21, 2012.

N. Mercer, S. Hennessy, and P. Warwick. Using interactive whiteboards to orchestrate classroom dialogue. *Technology, Pedagogy and Education*, 19:195–209, 2010a.

N. Mercer, P. Warwick, R. Kershner, and J. Kleine Staarman. Can the interactive whiteboard help to provide 'dialogic space' for children's collaborative activity? *Language and Education*, pages 367–384, 2010b.

N. Mercer, P. Warwick, and A. Ahmed. An oracy assessment toolkit: Linking research and development in the assessment of students' spoken language skills at age 11-12. *Learning and Instruction*, pages 51–60, 2017.

N. Mercer, R. Wegerif, and L. Major. *The Routledge international handbook of research on dialogic education.* Routledge, 2019.

R. Mitchell, F Myles, and E. Marsden. *Second Language Learning Theories*. Routledge, 2013.

Y. Okada, T. Sawaumi, and T. Ito. How do speech model proficiency and viewing order affect japanese efl learners' speaking performances. *Computer-Assisted Language Learning-Electronic Journal*, 19(2):61–81, 2018.

N. Pinkwart, C. Lynch, K. Ashley, and V. Aleven. Re-evaluating largo in the classroom: Are diagrams better than text for teaching argumentation skills? In *International Conference on Intelligent Tutoring Systems*, pages 90–100, Heidelberg, 2008. Springer.

S. Ruan, A. Willis, Q. Xu, G. M. Davis, L. Jiang, E. Brunskill, and J. A. Landay. Bookbuddy: Turning digital materials into interactive foreign language lessons through a voice chatbot. In *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*, pages 1–4, New York, 2019. ACM.

C. Sahil, R. Gianforte, and J. Sholar. Meet percy: The cs 221 teaching assistant chatbot. *ACM Transactions on Graphics*, 1(1):1–8, 2016.

K. Sedova. A case study of a transition to dialogic teaching as a process of gradual change. *Teaching and Teacher Education*, 67:278–290, 2017.

C. Stab and I. Gurevych. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1501–1510, Dublin, 2014. COLING.

I. M. Thies, N. Menon, S. Magapu, M. Subramony, and J. O'Neill. How do you want your chatbot? an exploratory wizard-of-oz study with young, urban indians. pages 441–459, Mumbai, 2017. IFIP Conference on Human-Computer Interaction.

M. West, R. Kraut, and H. E. Chew. I'd blush if i could: Closing gender divides in digital skills through education, 2019.

F. Yu, Z. Yao, X. Wang, K. An, L. Xie, Z. Ou, B. Liu, X. Li, and G. Miao. The slt 2021 children speech recognition challenge: open datasets, rules and baselines.

A Appendix I: dialogue acts

The following tables break down the potential interaction between the student and the dialogue system into all the dialogue acts considered. Jurafsky and Martin (2019) define a dialogue act as the function that an utterance performs in the dialogue, whether it is user input or a response from the system. Dialogue acts need to be correctly identified for the system to apply the most appropriate policy – deciding what the system needs to do (ibid). Some of the dialogue acts can be performed without Natural Language Processing, by having prewritten options that the user can choose from – those cases are indicated on the table with an asterisk after the dialogue act name. The majority of examples on the Example column are taken from the conversation in Appendix B.

DIALOGUE ACT	ACTOR	DESCRIPTION	EXAMPLE
ACKNOWLEDGING COMPREHENSION *	System/ User	Moving the conversation forward while indicating that the previous message was understood	Okay, seems easy enough ☺
AGREEING WITH THE AUTHOR *	User	Stating that their position on the issue of the task text is the same as the author's	Alright, let's say I kinda AGREE with the author for now
ASKING THE STUDENT TO MAKE A REBUTTAL	System	Presenting an idea to counter the student's latest argument and telling them to attempt a rebuttal	You make a good point, but what do you think about this? Many people argue that the gender pay gap is a "a feminist myth", so they would need to see reliable reports to see if it exists. Got a rebuttal against that?
ASKING THE STUDENT TO PRESENT AN ARGUMENT	System	Telling the student to defend their position with some reasons	Now tell me, which argument can you think of to defend your position?
ASKING THE STUDENT TO READ OUT LOUD	System	Telling the student to read the text	I'll show you the text paragraph by paragraph, and you simply have to read them. I'll only give you a new paragraph after I've heard you read the latest one. Give it your best! 🕮
ASKING THE STUDENT TO SELECT A PARAGRAPH	System	Telling the student to select a paragraph number for the paragraph that answers a particular question	Which paragraph contains the author's thesis? Select the paragraph number (if it appears in more than one, just pick one 😉)
ASKING THE STUDENT TO SUMMARIZE THE THESIS AND MAIN ARGUMENT	System	Telling the student to put the main ideas of the text into their own words	Could you put the ideas you just spotted into your own words? Basically: what is the author defending and which main argument do they use to back that?
CALLING OUT DISRESPECTFUL LANGUAGE	System	Calling the student's attention when they use inappropriate language – even if their answer was correct otherwise	Hey! ☹ Remember to speak respectfully! Otherwise you undermine your own arguments. You can surely do better. ☹
CONFIRMING *	User	Accepting a suggestion or saying "yes" to a question	Okay
DISAGREEING WITH THE AUTHOR *	User	Stating that their position on the issue of the task text is not the same as the author's	Well, let's say I DISAGREE with the author for now

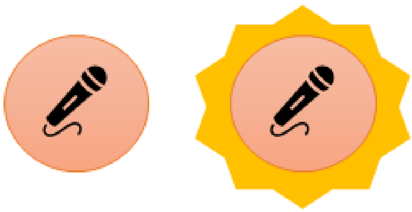
DIALOGUE ACT	ACTOR	DESCRIPTION	EXAMPLE
ENCOURAGING THE STUDENT	System	Giving positive feedback	Excellent! 😊
ERROR – NO INPUT	User	Making an utterance that cannot be detected by the system (e.g. opening the microphone by mistake, speaking too quietly, speaking too far from the microphone, etc.)	creak creak
ERROR – NO MATCH	User	Making an utterance that the system identifies as not related to the task (e.g. "trolling" the system, speaking to someone else with the microphone open, expressing task-related ideas in a very unusual way that the system was not trained to recognize, etc.)	Get away from the desk, kitty!
ERROR RESPONSE – NO INPUT	System	Assuming that the user did not hear the system's prompt and repeating it – but also trying not to be too repetitive	Mmm, I'm not hearing anything from you 👂 I'll give you the text again
ERROR RESPONSE – NO MATCH	System	Apologizing for the miscommunication and repeating the question that the user was answering	Sorry, I didn't understand that well. You were saying that the author's thesis is…?
EXPLAINING THE INTERFACE	System	Guiding the user on how to use the interface	If you don't know what to say, you can click on the lightbulb button to get some resources to inspire you
EXPLAINING THE TASK	System	Describing what is going to be done	First you'll read an argumentative text from the SAT. I'll be listening! 🙂
GIVING CORRECTIVE FEEDBACK	System	Telling the student how they might improve their answer	You seem to be on the right track 🙂 But I think you're focusing on the arguments and forgetting about the thesis. Remember all the paragraphs you selected on the previous exercise! Give it another try 💪
GREETING	System	Starting or ending the interaction with a friendly expression – the initiative is mainly taken by the system, so no greeting is required of the user	I hope I'll see you again soon!

DIALOGUE ACT	ACTOR	DESCRIPTION	EXAMPLE
MOVING TO ANOTHER PART OF THE APPLICATION *	User	Alternating between the dialogue, the text display and the diagram tool	Take me to the TEXT window first
PRESENTING A COUNTERARGUME NT	System/ User	Attacking another's argument	There already seems to be sufficient evidence of the gender gap in general to at least know it's real, so the people who still call it a myth don't do it for lack of evidence, but because they just don't wanna believe it.
PRESENTING AN ARGUMENT	User	Defending a position with some reasons	Publishing reports does nothing. There already is a global gender gap report and nobody cares about it. It's not like people don't know there's a gender gap, they just don't wanna do anything about it.
PROVIDING INFORMATION RESOURCES	System	Giving the student hints or links to websites that can help them come up with an answer	It's okay if you don't have a clear idea 🙂 Maybe taking a look at this link can help you make up you mind: https://en.wikipedia.org/wiki/Gender_pay_gap
RECITING A TEXT	System/ User	Reading the text that the task is based on	More than a half-century after President John F. Kennedy signed the Equal Pay Act of 1963...
REJECTING *	User	Saying "no" to a question or suggestion	Noo, I'm on fire! 😎 I have more arguments
REQUESTING CLARIFICATION *	System/ User	Asking for something to be explained	What do you mean by "THESIS"?
REQUESTING CONFIRMATION *	System	Asking the student a yes/no question or making a suggestion	Let me show you what we'll do
SELECTING A PARAGRAPH *	User	Selecting a paragraph number	5
SUGGESTING MOVING ON TO NEXT SUB-TASK	System	Acknowledging that the goals of current subtask have been met and letting the student continue with another sub-task	We've gathered quite a few arguments already, possibly enough to make a strong case ✨ Do you want to end it here and organize your thoughts into a graph?

B Appendix II: example conversation

Section 5.7 describes how the following example was designed. Nonetheless, the following images include a description of some elements of the system interface and how we have represented them on the still images.

The speech bubble styles and shapes are used throughout this appendix to mean:



This is the button to switch the microphone on and off. When the microphone is in, the button flashes (as represented by the yellow glow on the second version of the button)

The dialogue system is talking

This is a button the student can click on to select a response

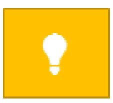
The dialogue system is giving the student a text to read

The student is talking

This is a drop-down menu where the student has selected the option being shown

The student has clicked on the button that this hand is pointing to

This is a button that the student can click on to return to the beginning of the argumentation



This is a button that the student can click on to get some help to come up with arguments

Hi, Timmy! I'm Robosan and I'm going to guide you through a task that will help you ace argumentation. But we'll go step by step, don't worry. Let me show you what we'll do.

Okay ☺

It's alright, I've already done this before

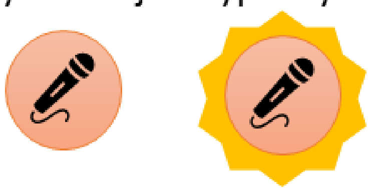
You sure? I just wanna make the task easy ☺

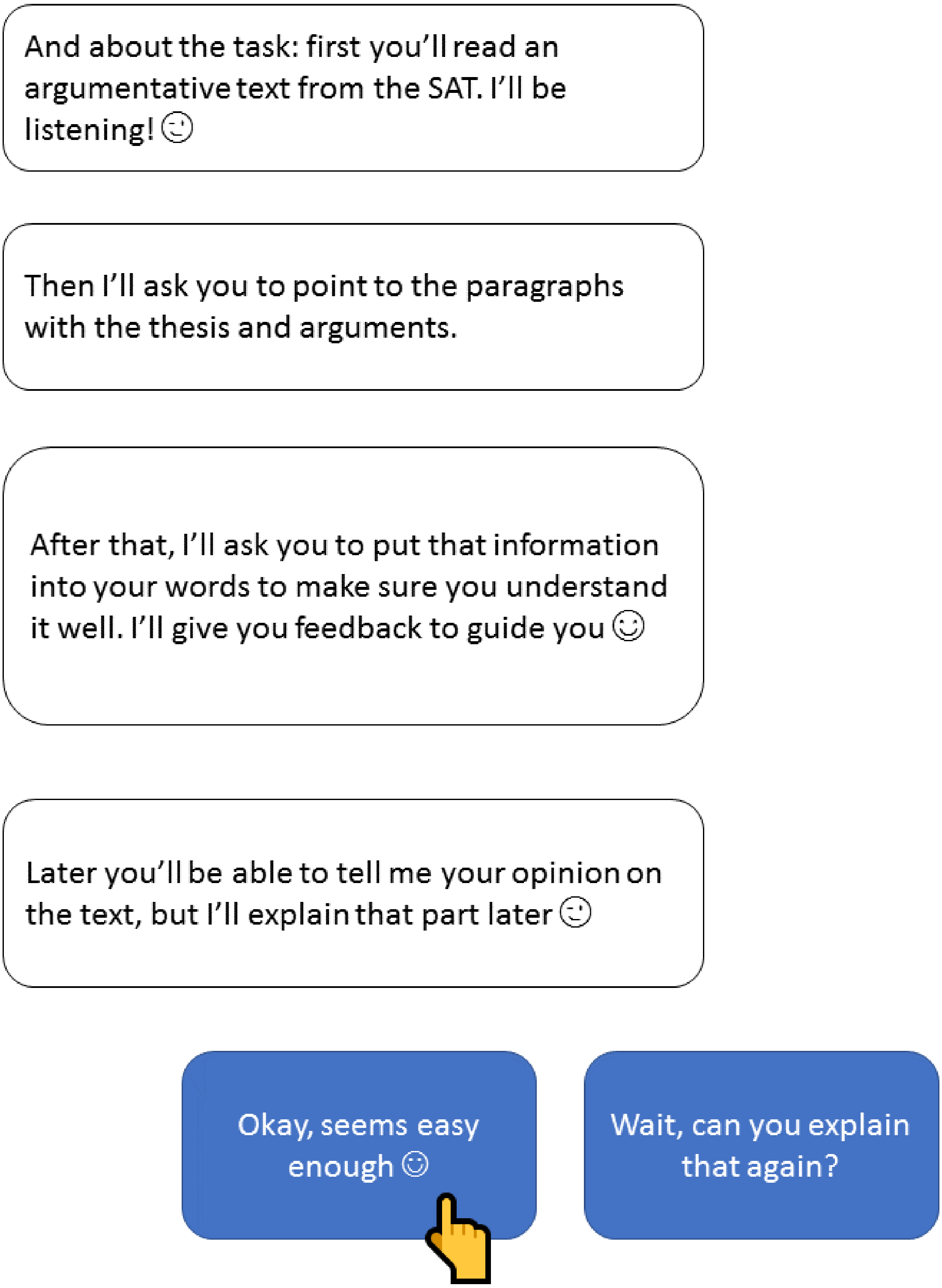
Ok, just explain it briefly

I pinky-promise I already did this before

Okay. First, the technical details: you'll sometimes need to answer by simply selecting from a couple of options, to make communication easier

Other times you'll be able to talk more normally, whenever you see a microphone button. Use it to turn the microphone on and off to talk to me...or you can just type if you prefer ☺




And about the task: first you'll read an argumentative text from the SAT. I'll be listening! ☺
Then I'll ask you to point to the paragraphs with the thesis and arguments.
After that, I'll ask you to put that information into your words to make sure you understand it well. I'll give you feedback to guide you ☺
Later you'll be able to tell me your opinion on the text, but I'll explain that part later ☺
Okay, seems easy enough ☺
Wait, can you explain that again?

Great! Let's begin. I'm going to give you an argumentative SAT text. You can access it anytime on the TEXT tab... but now let's read it together ☺ I gotta check that you don't skip the reading! 😇

Let's get reading ☺

Take me to the TEXT window first

Alright! I'll show you the text paragraph by paragraph, and you simply have to read them. I'll only give you a new paragraph after I've heard you read the latest one. Give it your best! 📖

HOW serious are we, really, about tackling income equality? . . .

HOW serious are we, really, about tackling income equality? . . .

More than a half-century after President John F. Kennedy signed the Equal Pay Act of 1963, the gap between what men and women earn has defied every effort to close it. And it can't be explained away as a statistical glitch, a function of women preferring lower-paying industries or choosing to take time off for kids.

More than a half-century after President John F. Kennedy signed the Equal Pay Act of 1963, the gap between what men and women earn has defied every effort to close it. And it can't be explained away as a statistical glitch, a function of women preferring lower-paying industries or choosing to take time off for kids.

Claudia Goldin, a labor economist at Harvard, has crunched the numbers and found that the gap persists for identical jobs, even after controlling for hours, education, race and age. Female doctors and surgeons, for example, earn 71 percent of what their male colleagues make, while female financial specialists are paid just 66 percent as much as comparable men. Other researchers have calculated that women one year out of college earn 6.6 percent less than men after controlling for occupation and hours, and that female M.B.A. graduates earn on average $4,600 less than their male classmates for their first jobs.

Claudia Goldin, a labor economist at Harvard, has crunched the numbers and found that the gap persists for identical jobs, even after controlling for hours, education, race and age. Female doctors and surgeons, for example, earn 71 percent of what their male colleagues make, while female financial specialists are paid just 66 percent as much as comparable men. Other researchers have calculated that women one year out of college earn 6.6 percent less than men after controlling for occupation and hours, and that female M.B.A. graduates earn on average $4,600 less than their male classmates for their first jobs.

It's not that men are intentionally discriminating against women—far from it. I've spent the past year interviewing male executives for a book about men and women in the workplace. A vast majority of them are fair-minded guys who want women to succeed. They're absolutely certain that they don't have a gender problem themselves; it must be some other guys who do. Yet they're leaders of companies that pay men more than women for the same jobs.

It's not that men are intentionally discriminating against women—far from it. I've spent the past year interviewing male executives for a book about men and women in the workplace. A vast majority of them are fair-minded guys who want women to succeed. They're absolutely certain that they don't have a gender problem themselves; it must be some other guys who do. Yet they're leaders of companies that pay men more than women for the same jobs.

Women are trying mightily to close that chasm on their own. Linda Babcock, an economist at Carnegie Mellon and co-author of the book "Women Don't Ask," has found that one reason for the disparity is that men are four times more likely to ask for a raise than women are, and that when women do ask, we ask for 30 percent less. And so women are told we need to lean in, to demand to be paid what we're worth. It's excellent advice—except it isn't enough.

Women are trying mightily to close that chasm on their own. Linda Babcock, an economist at Carnegie Mellon and co-author of the book "Women Don't Ask," has found that one reason for the disparity is that men are four times more likely to ask for a raise than women are, and that when women do ask, we ask for 30 percent less. And so women are told we need to lean in, to demand to be paid what we're worth. It's excellent advice—except it isn't enough.

There is an antidote to the problem. Britain recently introduced a plan requiring companies with 250 employees or more to publicly report their own gender pay gap. It joins a handful of other countries, including Austria and Belgium, that have introduced similar rules. (In the United States, President Obama last year signed a presidential memorandum instructing federal contractors to report wage information by gender and race to the Department of Labor.) The disclosures "will cast sunlight on the discrepancies and create the pressure we need for change, driving women's wages up," Prime Minister David Cameron said last month.

There is an antidote to the problem. Britain recently introduced a plan requiring companies with 250 employees or more to publicly report their own gender pay gap. It joins a handful of other countries, including Austria and Belgium, that have introduced similar rules. (In the United States, President Obama last year signed a presidential memorandum instructing federal contractors to report wage information by gender and race to the Department of Labor.) The disclosures "will cast sunlight on the discrepancies and create the pressure we need for change, driving women's wages up," Prime Minister David Cameron said last month.

Critics of the British plan protest that it's too expensive and complex. Some contend that it doesn't address the root of the problem: systemic issues that block women from higher-paying industries, and social issues like unconscious bias.

Critics of the British plan protest that it's too expensive and complex. Some contend that it doesn't address the root of the problem: systemic issues that block women from higher-paying industries, and social issues like unconscious bias.

But real-world results suggest otherwise. Last year, the consulting firm PricewaterhouseCoopers voluntarily released its gender pay gap in Britain, one of five firms in the country, including AstraZeneca, to do so. Simply saying the number out loud "created much more momentum internally" to close it, Sarah Churchman, who runs the firm's British diversity and inclusion efforts, told me.

But real-world results suggest otherwise. Last year, the consulting firm PricewaterhouseCoopers voluntarily released its gender pay gap in Britain, one of five firms in the country, including AstraZeneca, to do so. Simply saying the number out loud "created much more momentum internally" to close it, Sarah Churchman, who runs the firm's British diversity and inclusion efforts, told me.

PricewaterhouseCoopers's analysis showed that most of its 15.1 percent pay disparity (compared with a Britain-wide gap of more than 19 percent) reflected a lack of women in senior jobs. So the firm focused on whether it was promoting fairly. In 2013, the grade just below partner was 30 percent female, yet only 16 percent of those promoted to partner were women. A year later, the percentage of women promoted to partner had more than doubled..

PricewaterhouseCoopers's analysis showed that most of its 15.1 percent pay disparity (compared with a Britain-wide gap of more than 19 percent) reflected a lack of women in senior jobs. So the firm focused on whether it was promoting fairly. In 2013, the grade just below partner was 30 percent female, yet only 16 percent of those promoted to partner were women. A year later, the percentage of women promoted to partner had more than doubled..

The potential cost savings of publishing the gender wage gap are enormous. About 20 percent of large companies now train employees to recognize unconscious bias, spending billions of dollars to try to stamp out unintentional discrimination. Paying for a salary analysis is cheaper and potentially more effective. Evidence also suggests that less secrecy about pay results in greater employee loyalty and lower turnover. . . .

The potential cost savings of publishing the gender wage gap are enormous. About 20 percent of large companies now train employees to recognize unconscious bias, spending billions of dollars to try to stamp out unintentional discrimination. Paying for a salary analysis is cheaper and potentially more effective. Evidence also suggests that less secrecy about pay results in greater employee loyalty and lower turnover. . . .

Political realities being what they are, the chances of achieving [full] transparency are slim; even the tepid C.E.O. pay gap rule took the S.E.C. five years to push through, in the face of fierce industry opposition.

Creak creak...

Mmm, I'm not hearing anything from you I'll give you the text again.

Political realities being what they are, the chances of achieving [full] transparency are slim; even the tepid C.E.O. pay gap rule took the S.E.C. five years to push through, in the face of fierce industry opposition.

Political realities being what they are, the chances of achieving [full] transparency are slim; even the tepid C.E.O. pay gap rule took the S.E.C. five years to push through, in the face of fierce industry opposition.

But why would we not want a measure that will settle the controversy over the pay gap with quantifiable facts? Shining some much-needed sunlight on the gender wage gap will make a difference for every one of us, men and women, right now.

But why would we not want a measure that will settle the controversy over the pay gap with quantifiable facts? Shining some much-needed sunlight on the gender wage gap will make a difference for every one of us, men and women, right now.

Good job! 😆 Now take a look at the text ... Which paragraph contains the author's thesis? Select the paragraph number (if it appears in more than one, just pick one 😉)

What do you mean by "THESIS"?

-

The thesis is where the author states their position on a topic - it can also be called their main claim. Maybe they're defending an idea, or they are advocating for a specific course of action. Try to find that in the text.

5

Are you sure? Maybe there's a bit of the thesis there, but I think this paragraph is more about **describing the issue** [Despite good intentions from men and efforts by women, there is a gender gap in salaries]. Look for something more explicit ☺

6

Well spotted! 😄 And how does the author defend this position? If they emphasize one argument over others, select that one. Otherwise, pick any paragraph where they present their arguments.

8

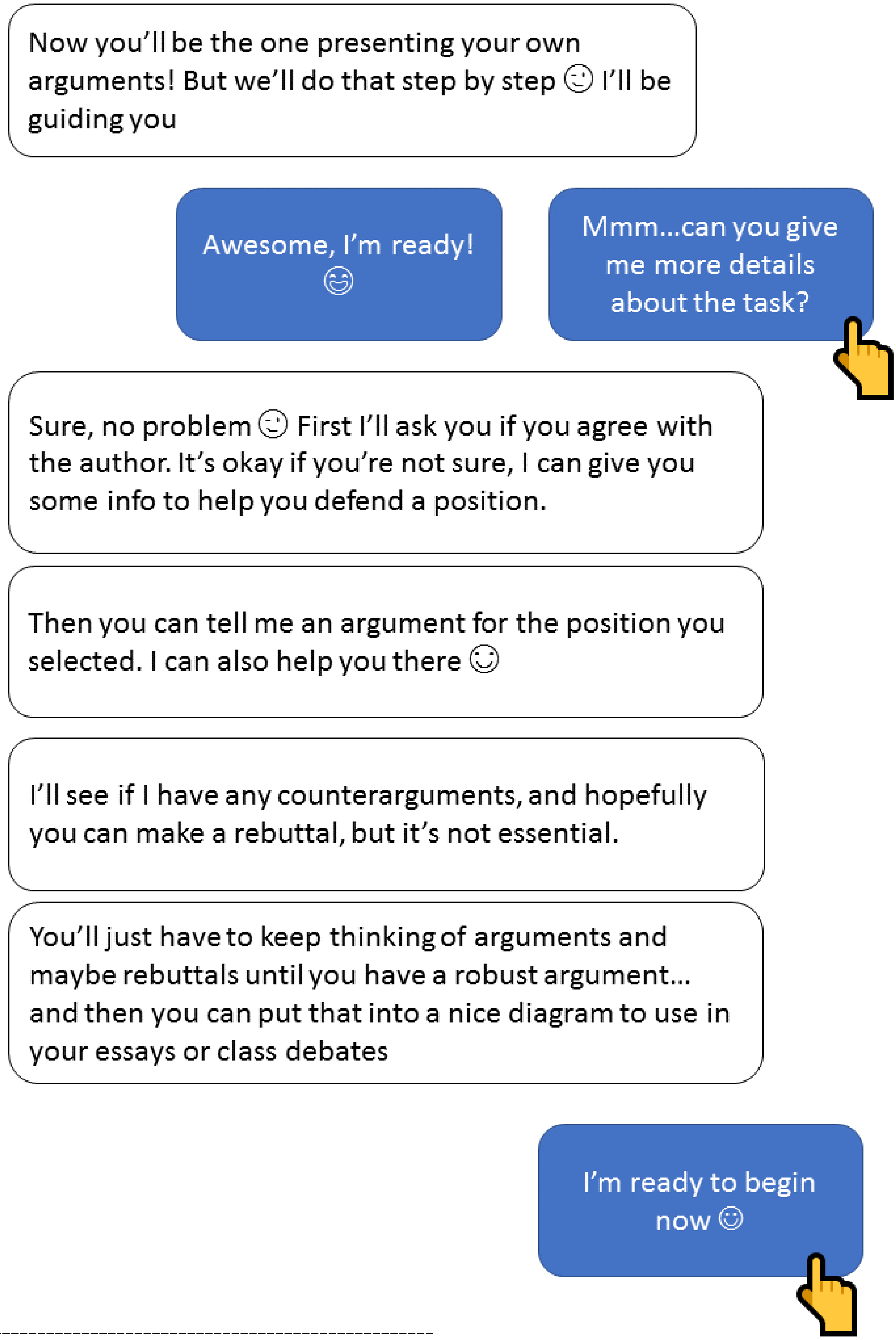
Good job! 👍 Could you put that info in your own words? Basically: **what is the author defending** and **which main argument** do they use to back that?

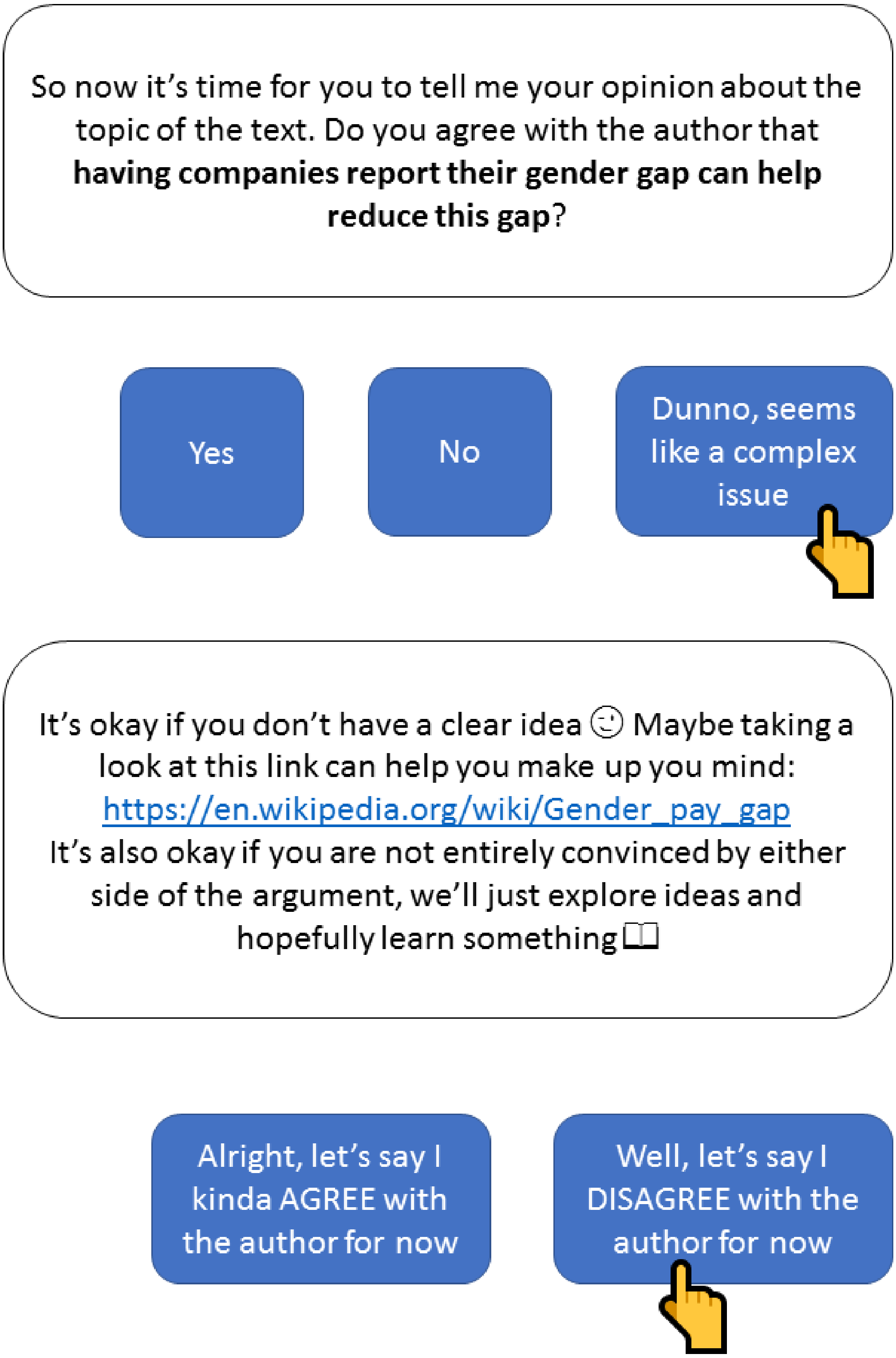
companies that know they have a gender gap feel more pressure to end it

You seem to be on the right track 🙂 But I think you're focusing on the arguments and forgetting about the **thesis.** Remember all the paragraphs you selected on the previous exercise! Give it another try 💪

Making companies say how much they pay men and women will raise awareness about sexist policies

Excellent! 😊 I think you're ready for the next exercise 🎉


Now you'll be the one presenting your own arguments! But we'll do that step by step I'll be guiding you
Awesome, I'm ready!
Mmm...can you give me more details about the task?
Sure, no problem First I'll ask you if you agree with the author. It's okay if you're not sure, I can give you some info to help you defend a position.
Then you can tell me an argument for the position you selected. I can also help you there
I'll see if I have any counterarguments, and hopefully you can make a rebuttal, but it's not essential.
You'll just have to keep thinking of arguments and maybe rebuttals until you have a robust argument... and then you can put that into a nice diagram to use in your essays or class debates
I'm ready to begin now


So now it's time for you to tell me your opinion about the topic of the text. Do you agree with the author that **having companies report their gender gap can help reduce this gap**?
Yes
No
Dunno, seems like a complex issue
It's okay if you don't have a clear idea 🙂 Maybe taking a look at this link can help you make up you mind: https://en.wikipedia.org/wiki/Gender_pay_gap
It's also okay if you are not entirely convinced by either side of the argument, we'll just explore ideas and hopefully learn something 📖
Alright, let's say I kinda AGREE with the author for now
Well, let's say I DISAGREE with the author for now

Good. Now tell me, which argument can you think of to defend your position?

- If going through the ideas you change your mind, you can click on the return button and restart your argumentation
- If you don't know what to say, you can click on the lightbulb button to get some resources to inspire you

Publishing reports does nothing. There already is a global gender gap report and nobody cares about it. It's not like people don't know there's a gender gap, they just don't wanna do anything about it.

You make a good point, but what do you think about this?

Many people argue that the gender pay gap is a "a feminist myth", so they would need to see reliable reports to see if it exists.

Got a rebuttal against that?

Yes

No, I got nothing to respond to that

Go ahead then! ☺

There already seems to be sufficient evidence of the gender gap in general to at least know it's real, so the people who still call it a myth don't do it for lack of evidence, but because they just don't wanna believe it.

Fair enough. Got another argument to make your case stronger? Tell me ☺

What do you think of the arguments that the author used? If you're defending the opposite view, maybe you can defend your position with a rebuttal of some of those arguments.

Well, the author mentions some companies where reporting the gender gap gave good results, but it's probably because there were already good intentions in those companies. If the gender gap exists it is because those good intentions are pretty rare in society, so giving people more data about how women are paid less won't work most of the time.

Alright, that's an interesting point. But I have a counterargument for that:

Women account for around half the population, so they could account for around half of consumers. Even if many companies don't care much about their sexist practices, they'll probably care about their reputation, as a bad reputation could translate in fewer customers.

Can you think of a rebuttal for that?

Yes

No, I got nothing to respond to that

It's okay. But maybe you can think of another argument for your main claim? Give it a try!

Mmmm, the gender gap exists because the dickheads in power want to stay in power and give nothing to women. They're just gonna wipe their ass with all the reports, cos they have the power to ignore women.

Hey! ☹ Remember to speak respectfully! Otherwise you undermine your own arguments. You can surely do better. ☹ Try again.

Ok...The gender gap exists because men are in power and want to stay in power. Reports are not powerful enough to change the way the world has worked for ages.

Okay...You've left me speechless. I have no counterargument against that.

We've gathered quite a few arguments already, possibly enough to make a strong case ✨ Do you want to end it here and organize your thoughts into a graph?

Yeah, let's do the GRAPH

Noo, I'm on fire! 😎 I have more arguments

Alright, I'll take you to the GRAPH tool. There you'll have everything you need The window has a help function if you want me to show you how the graph tool works or give you tips on how to build your graph ☺

Btw, it's been a pleasure talking to you. I hope I'll see you again soon! 😄

To the GRAPH window

C Appendix III: example annotated text

The following is the May 2019 SAT text with our annotations. The first column is the paragraph number, and the second one contains the paragraphs. Then, the third column contains the label assigned to the paragraph, followed by a summary on the fourth column. The remaining texts, together with the answer data, can be accessed at: Link to SAT dataset.

paragraph_number	text	function1	summary_of_function1
1	HOW serious are we, really, about tackling income equality? . . .	describing the issue	Despite good intentions from men and efforts by women, there is a gender gap in salaries
2	More than a half-century after President John F. Kennedy signed the Equal Pay Act of 1963, the gap between what men and women earn has defied every effort to close it. And it can't be explained away as a statistical glitch, a function of women preferring lower-paying industries or choosing to take time off for kids	describing the issue	Despite good intentions from men and efforts by women, there is a gender gap in salaries
3	Claudia Goldin, a labor economist at Harvard, has crunched the numbers and found that the gap persists for identical jobs, even after controlling for hours, education, race and age. Female doctors and surgeons, for example, earn 71 percent of what their male colleagues make, while female financial specialists are paid just 66 percent as much as comparable men. Other researchers have calculated that women one year out of college earn 6.6 percent less than men after controlling for occupation and hours, and that female M.B.A. graduates earn on average $4,600 less than their male classmates for their first jobs.	describing the issue	Despite good intentions from men and efforts by women, there is a gender gap in salaries
4	It's not that men are intentionally discriminating against women—far from it. I've spent the past year interviewing male executives for a book about men and women in the workplace. A vast majority of them are fair-minded guys who want women to succeed. They're absolutely certain that they don't have a gender problem themselves; it must be some other guys who do. Yet they're leaders of companies that pay men more than women for the same jobs.	describing the issue	Despite good intentions from men and efforts by women, there is a gender gap in salaries
5	Women are trying mightily to close that chasm on their own. Linda Babcock, an economist at Carnegie Mellon and co-author of the book "Women Don't Ask," has found that one reason for the disparity is that men are four times more likely to ask for a raise than women are, and that when women do ask, we ask for 30 percent less. And so women are told we need to lean in, to demand to be paid what we're worth. It's excellent advice—except it isn't enough.	describing the issue	Despite good intentions from men and efforts by women, there is a gender gap in salaries
6	There is an antidote to the problem. Britain recently introduced a plan requiring companies with 250 employees or more to publicly report their own gender pay gap. It joins a handful of other countries, including Austria and Belgium, that have introduced similar rules. (In the United States, President Obama last year signed a presidential memorandum instructing federal contractors to report wage information by gender and race to the Department of Labor.) The disclosures "will cast sunlight on the discrepancies and create the pressure we need for change, driving women's wages up," Prime Minister David Cameron said last month.	thesis	Having companies report their gender gap can help reduce this gap
7	Critics of the British plan protest that it's too expensive and complex. Some contend that it doesn't address the root of the problem: systemic issues that block women from higher-paying industries, and social issues like unconscious bias.	counterargument not rebutted in the same paragraph	Having companies report their gender gap is too costly and doesn't solve the issue
8	But real-world results suggest otherwise. Last year, the consulting firm PricewaterhouseCoopers voluntarily released its gender pay gap in Britain, one of five firms in the country, including AstraZeneca, to do so. Simply saying the number out loud "created much more momentum internally" to close it, Sarah Churchman, who runs the firm's British diversity and inclusion efforts, told me.	argument	In an example company where they report their gender gap, this has created awareness and a bigger desire to solve the problem
9	PricewaterhouseCoopers's analysis showed that most of its 15.1 percent pay disparity (compared with a Britain-wide gap of more than 19 percent) reflected a lack of women in senior jobs. So the firm focused on whether it was promoting fairly. In 2013, the grade just below partner was 30 percent female, yet only 16 percent of those promoted to partner were women. A year later, the percentage of women promoted to partner had more than doubled.	argument	In an example company where they report their gender gap, this has allowed them to find the root of the problem
10	The potential cost savings of publishing the gender wage gap are enormous. About 20 percent of large companies now train employees to recognize unconscious bias, spending billions of dollars to try to stamp out unintentional discrimination. Paying for a salary analysis is cheaper and potentially more effective. Evidence also suggests that less secrecy about pay results in greater employee loyalty and lower turnover. . . .	argument	Reporting the gender gap is cheaper than other strategies to tackle this issue
11	Political realities being what they are, the chances of achieving [full] transparency are slim; even the tepid C.E.O. pay gap rule took the S.E.C. five years to push through, in the face of fierce industry opposition.	counterargument not rebutted in the same paragraph	Passing and enforcing legislation to have companies report their gender gap will be difficult
12	But why would we not want a measure that will settle the controversy over the pay gap with quantifiable facts? Shining some much-needed sunlight on the gender wage gap will make a difference for every one of us, men and women, right now.	argument	Having companies report their gender gap will make the scope of this controversial issue clearer