Teaching Students About Conversational AI Using CONVO, a Conversational Programming Agent

Jessica Zhu
MIT CSAIL
Cambridge, USA
jfzhu@mit.edu

Jessica Van Brummelen
MIT CSAIL
Cambridge, USA
jess@csail.mit.edu

***Abstract*—Smart assistants, like Amazon's Alexa or Apple's Siri, have become commonplace in many people's lives, appearing in their phones and homes. Despite their ubiquity, these conversational AI agents still largely remain a mystery to many, in terms of how they work and what they can do. To lower the barrier to entry to understanding and creating these agents for young students, we expanded on CONVO, a conversational programming agent that can respond to both voice and text inputs. The previous version of CONVO focused on teaching only programming skills, so we created a simple, intuitive user interface for students to use those programming skills to train and create their own conversational AI agents. We also developed a curriculum to teach students about key concepts in AI and conversational AI in particular. We ran a 3-day workshop with 15 participating middle school students. Through the data collected from the pre- and post-workshop surveys as well as a mid-workshop brainstorming session, we found that after the workshop, students tended to think that conversational AI agents were less intelligent than originally perceived, gained confidence in their abilities to build these agents, and learned some key technical concepts about conversational AI as a whole. Based on these results, we are optimistic about CONVO's ability to teach and empower students to develop conversational AI agents in an intuitive way.**

***Index Terms*—conversational AI, conversational AI agent, education, humman-computer interaction, intent, unconstrained natural language, conversational programming**

I. INTRODUCTION AND RELATED WORK

With conspiracy theories and misunderstandings about how conversational agents (CAs) work [1]–[3], and how quickly CAs, like Alexa and Siri, have become household names [4], [5], AI and conversational AI education is becoming increasingly important [6]–[8]. Furthermore, conversational AI technology has become increasingly useful in educational contexts. For example, researchers have designed agents to help students manage emotions during learning, teach history, and quiz students [9]–[11]. Other agents, like Betty's Brain and Zhorai, are teachable themselves, drawing on the learning-by-teaching paradigm [12]–[14]. Still others draw on CAs' abilities to lower the barrier to entry for people to develop skills, like programming [15].

Despite the need for AI and CA education, and the evident utility CAs can provide, high-utility CA development interfaces, like 'Actions on Google', which could be useful for CA education, often have steep learning curves [16]–[18]. Furthermore, current low-barrier-to-entry CA development interfaces, like 'Alexa Blueprints', generally lack many of the features high-utility interfaces include [16], [19]. These low-barrier-to-entry interfaces are generally not developed to educate people about how CAs work either. Nonetheless, one CA interface with this purpose includes Conversational AI in MIT App Inventor, which has been used in K-12 settings to teach students about AI as they develop CAs [20]. This interface has been shown to be an effective tool in teaching AI literacy concepts and students to program; however, the programming itself is done through a visual code-block interface, rather than a speech- or conversational-based interface [8]. We posit that by utilizing conversational AI to program (e.g., having students *tell* a system to "create a procedure"), students could learn further about CAs and how AI works during the development process itself. For instance, students could learn about the suitability of constrained (e.g., regular expression matching) versus unconstrained (e.g., deep learning classification) natural language (NL) systems through using both systems when conversing with the CA [21], [22].

Recently developed interfaces that utilize conversational programming include SUGILITE, which is an agent that automates tasks through NL and GUI interactions, and CONVO, which is an open-source agent that develops Python-based programs through NL conversation [15], [23]–[25]. SUGILITE allows users to demonstrate a GUI-based task and describe the task in words. The multimodal inputs (GUI and speech interactions) increase SUGILITE's performance and allow users without significant programming experience to automate tasks [23]. CONVO instead uses purely NL input (e.g., speech) to allow users to program conversation-based tasks. For example, a user may program CONVO to play a 20-questions game by conversing with it in NL [15].

In this study, we investigate how CONVO could benefit from GUI interactions (similarly to SUGILITE and MIT App Inventor) and furthermore be used as a teaching tool for conversational AI in addition to computer programming. We do so through developing a GUI for users to define NL intents and train ML models, and engaging 15 middle school students in 3-day, remote (Zoom-based) workshops. We also investigate students' self-efficacy and perceptions of CAs before and after the workshops, as these results could enable us to develop CAs

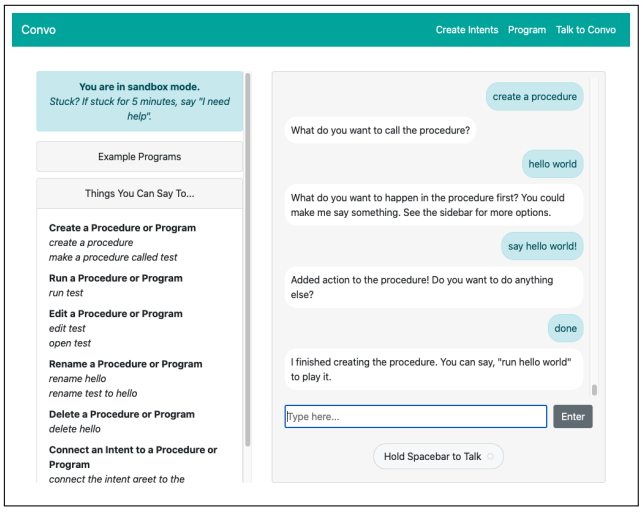

Fig. 1. The new user interface of CONVO. It maintains all of the same functionality of the previous version of CONVO, including the possible commands CONVO understands, as well as the voice and text inputs.

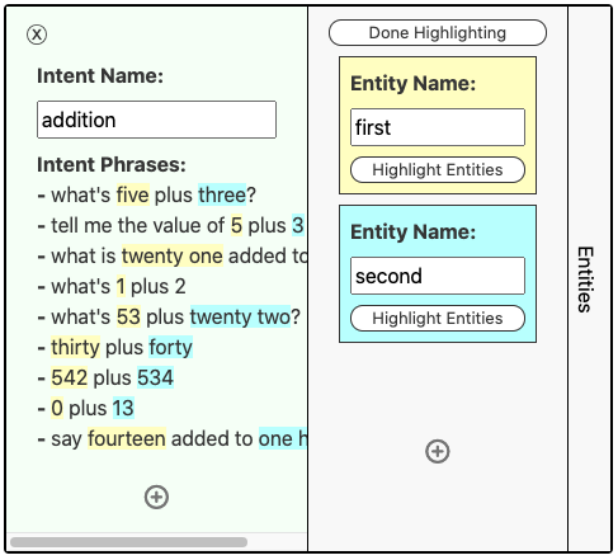

Fig. 2. An example intent card that contains the training data a user might input to teach CONVO how to recognize when someone wants to add two numbers together. The intent phrases are example inputs a user might enter to trigger this intent, and the two entities (the *first* and *second* numbers to add together) are the two pieces of information CONVO needs to learn to extract from an intent phrase. Because of Rasa and BERT, with just a few training examples, CONVO is able to generalize across NL, recognizing phrases that are not necessarily in the given intent phrases (e.g., "give me nine plus two").

that better address students' learning needs [26]–[28].

Through the data collected in the workshops where students used CONVO to train ML models and create conversational apps, we aimed to answer the following research questions:

RQ1: What conversational AI and AI literacy concepts can students learn through our CA workshops?

RQ2: How do students' feelings of self-efficacy and perceptions of CAs change through our workshops?

We posit that the data from the study will benefit the greater CA and AI education research communities through CONVO's novel GUI system design, the five key takeaways from this study, and the opportunity for extended follow-up studies.

II. TECHNICAL IMPLEMENTATION

To allow CONVO to support unconstrained NL CA-creation, we completed three main tasks. The first was to provide CONVO with a less constrained way to interpret user utterances, such that a user did not need to use an exact syntax pattern when communicating with CONVO [15]. To do so, we utilized Rasa, a machine learning (ML) framework, to create ML models while relying on BERT, a pre-trained NL understanding (NLU) model [29], [30]. Rasa is responsible for the tasks of "learning" about the training data, recognizing user intent, and extracting entities from user input.

After integrating Rasa with CONVO, we created a new user interface for CONVO, part of which can be seen in Fig. 1 to provide an intuitive way for students to enter in training data for CONVO to learn. The training data includes intent phrases (groups of sentences with similar meanings) and entity examples (specific information in intents; e.g., a date, like "May 2") [31]. A user teaching CONVO to recognize someone wanting to add two numbers together might fill out a card with the data seen in Fig. 2. Once users finish inputting data, they would be able to click a button to start the ML training process.

After developing a way to train CONVO to recognize intents and extract necessary information through entities, we combined this capability with CONVO's existing conversational programming abilities [15]. To do so, we created a new user flow to connect intents with procedures, in which users would speak or type "connect the intent `<intent name>` to the procedure `<procedure name>`". Once an intent was connected to a procedure, any time that intent was recognized, CONVO would run the corresponding procedure. For example, a user might train CONVO to recognize the intent, "say hello", program a procedure in which CONVO speaks, "Hello World!", and then connect the intent to the corresponding procedure.

III. WORKSHOP

To test our prepared curriculum and CONVO, we held a series of three 2-hour long workshops over consecutive Saturdays. The workshops were run through SPARK, a student-run MIT initiative for teaching 7th and 8th graders [32].

Our curriculum was partially based on resources from an open-source MIT App Inventor conversational AI workshop and can be found online [26], [33]. We used the same Big AI Ideas as a framework for explaining AI and conversational AI concepts [7]. In the pre- and post-workshop surveys, we asked some of the same questions to assess students' AI literacy [8].

New materials and ideas we introduced included applying the Big AI Ideas to conversational AI specifically, discussing the spectrum of unconstrained and constrained NL, and teaching students to develop CAs with CONVO.

IV. RESULTS

Students ranged from 11 to 14 years old, and were in 7th or 8th grade. We obtained data for 12 students (4 female, 8 male)

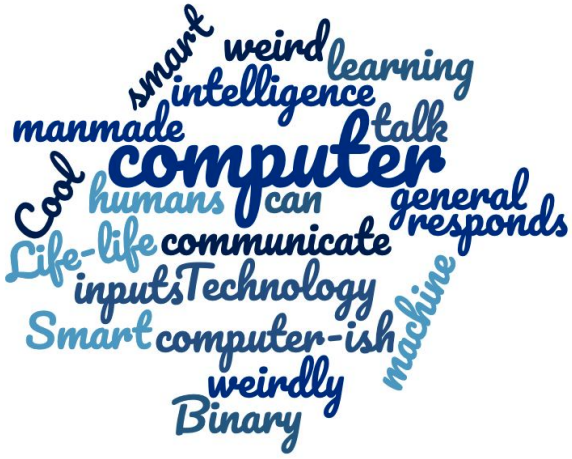

Fig. 3. A graphic of the words students used to describe AI in the pre-workshop survey. Larger words correspond to a larger word frequency. Figure generated using [34].

in the pre-workshop survey and 7 in the post-workshop survey. Ten out of 12 students had some form of prior programming experience, either with block-based (e.g. Scratch, MIT App Inventor) or text-based (e.g. Python, Java) programming, but this was not a pre-requisite for the workshop.

A. Student Literacy

In the pre-workshop survey, we asked students to describe AI in just 3 words or short phrases. The aggregation of the 12 responses is found in Fig. 3. Common themes among the words and phrases were that computers and technology were smart and intelligent. Some students commented on how "weird" it was, and others gave responses that indicated that AI was still very much a black box topic to them.

Next, when students were asked to briefly describe how they thought CAs worked, only 5 out of 12 students responded with something related to code and/or programming. Even for those responses, they were often very vague. Other responses (5) mentioned responding to human inputs and one response simply stated "I don't know."

After the workshop, we asked students to describe what they thought CAs were. A total of 7 students gave responses, and no student said something that was inaccurate. Three students specifically stated you must program the agents with human-given data, and all seven students mentioned that these agents talk back to you. Overall, the students gave much more technical responses, showing their understanding of how CAs (and AI in general) rely heavily on humans for data and input.

They were also asked the same question as in the pre-workshop survey of how they thought CAs worked. Student responses to this question were much more technical here too, with many students using new terminology like "training," "data," and "entities." While some answers were still vague, no answer was incorrect, and every student mentioned needing a human to code or program something for the agent to respond.

Additionally, in the post-workshop survey, students were asked a series of true/false questions regarding conversational AI concepts in the context of CONVO. Results are shown in Table I. Students performed very well on the first two questions but poorly on the last three. The first two questions were related to the idea that more training data means a more accurate resulting agent while the last three questions surrounded the concept of constrained and unconstrained NL models. While students seemed to understand the first idea, it appears students had trouble fully grasping the latter idea.

TABLE I
STUDENTS WERE ASKED TO SELECT THE TRUE STATEMENTS FROM A SERIES OF STATEMENTS ABOUT CONVO. THESE STATEMENTS TARGETED SPECIFIC CONVERSATIONAL AI CONCEPTS.

Statement	**Correct (out of 7)**
When training an intent, it is better to have fewer training examples. (False)	7
When setting an entity, it is better to have more training examples. (True)	6
In the Program mode, CONVO would understand it when you say something different but similar to a command in the sidebar, for example replacing the word "procedure" with "function". (False)	3
In the Talk to Me mode, CONVO would understand it when you say something different but similar to a command in the sidebar, for example replacing the word "procedure" with "function". (True)	2
CONVO can recognize intents in both the Program and Talk to Me modes. (False)	3

B. Student Self-efficacy and Perceptions

Next, we observe how student confidence, interest, and perceptions of CAs changed through our workshop. We collected data on students' attitudes and how they might characterize a CA through the pre- and post-workshop surveys. We asked students to indicate how confident and interested they were in creating a CA using a 7-point Likert scale, where a point value of 1 corresponded to either *not at all confident* or *not at all interested* and a point value of 7 corresponded to either *extremely confident* or *extremely interested*. To substantiate our results, we used the Wilcoxon Signed-Rank Test to measure the magnitude of change between the pre- and post-workshop survey data [35].

From the results shown in Fig. 4, we see that at the end of the workshop, students generally felt much more confident in their abilities to create their own CAs ($\bar{x}_{diff} = 2.0$, $Md_{diff} = 2$, $|Z| = 2.03$, $p = 0.021$). Of the 7 students who also filled out both surveys, their confidence levels all either remained the same or increased.

As for student interest in CAs, the post-workshop survey results do not provide evidence for a significant difference before and after the workshop ($\bar{x}_{diff} = 0.0$, $Md_{diff} = 0$, $p > 0.05$). This is not too unexpected, given that student interest started off very high. Since student interest remained high even after the workshop, we are not concerned with the ability of the curriculum and CONVO to evoke interest.

To investigate student perceptions, we asked students to answer a series of *Persona Questions* as outlined in a previously

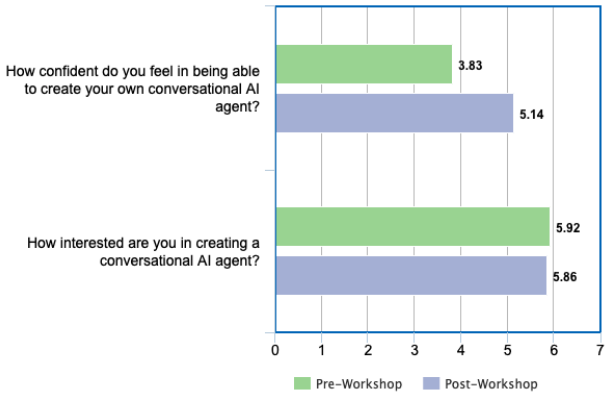

Fig. 4. Student attitudes based on a 7-point Likert scale towards CAs in the pre- and post-workshop surveys.

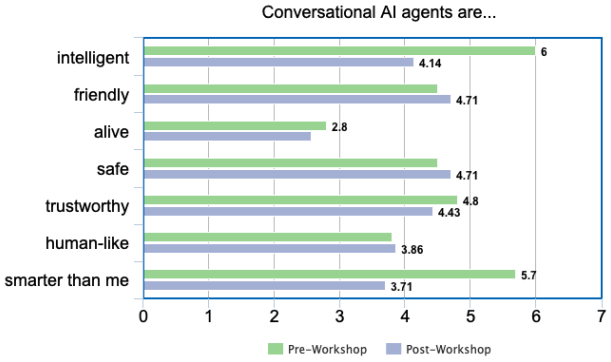

Fig. 5. Students rank how much they agree or disagree with statements describing CAs on a 7-point Likert scale in pre- and post-workshop surveys.

mentioned CA study [26]. These questions asked students to rank a series of statements about CAs on a 7-point Likert scale.

The results shown in Fig. 5 indicate that on average, after the workshop, students felt more or less neutral about every statement, with a slight disagreement with the safeness of CAs. We observe large changes, however, with the statements that CAs are intelligent ($\bar{x}_{diff} = -1.9$, $Md = -2$, $|Z| = 2.15$, $p = 0.016$) and smarter than them ($\bar{x}_{diff} = -2.0$, $Md = -2$, $|Z| = 2.21$, $p = 0.013$). We attribute this shift in sentiment to how students had learned about how much human work goes into making a CA seem 'smart.' For example, when working with CONVO, students had to input data, train a model, and tell CONVO exactly what to do in very specific scenarios.

Nevertheless, these results oppose the results in the original CA study with the *Persona Questions*, in which students ranked the CA, Amazon Alexa, as *more* intelligent after learning how to program it [26]. This may be due to different CAs being used the studies or different age ranges, and future studies may investigate this difference further.

C. Key Takeaways

- **Learning about and using CONVO empowered students to be more confident in their abilities to create their own CAs.** Student responses on the pre- and post-workshop surveys indicated a difference in how confident they were in their abilities to create their own CAs.
- **Students' perceptions of AI's intelligence shifted.** At the start of the workshop, students overwhelmingly agreed with the sentiments of CAs being 'intelligent' and 'smarter than them'. However, by the end of the workshop, students' opinions had changed, with students generally disagreeing with the same two sentences.
- **Students were able to learn some key concepts about conversational AI.** From class discussion and also the post-workshop survey results, it is clear that students were able to gain knowledge about the Big Five AI ideas. Students also showed improved knowledge in the areas of providing training data to agents and the steps required to create an agent, but failed to fully grasp the concept of constrained vs. unconstrained NL.
- **Students were able to create their own CAs.** Almost all students were able to complete the tutorials, which involved creating two separate CAs. Some students were able to venture even farther, and create original CAs.
- **CONVO is a useful tool that can act as a starting point for students to learn more about conversational AI and CAs in particular.** CAs are quite complex, and creating them is often a very involved and obscure process. Through this workshop, middle school students were able to learn about and create CAs, a promising step to empowering all students of any age or background to do so as well.

V. LIMITATIONS

One limitation of our study is the small sample size (7) of student data. We would require another, larger study to be performed to achieve higher confidence in our conclusions. Additionally, our workshop only targeted 7th and 8th graders who self-selected to participate, so our results may not generalize to all students. We encourage additional studies that span a larger range of ages and initial interest levels in CAs.

It is also possible that differences in environment and/or timing of the pre- and post-workshop surveys contributed to biases in our data. Especially because the workshop was held remotely, it was also sometimes difficult to determine when students' level of understanding. This discrepancy might have led to students not fully understand the concept of constrained and unconstrained NL, so we propose future studies to include more checkpoints around this topic.

VI. CONCLUSION AND FUTURE WORK

We conclude that the current iteration of CONVO has shown promising results for its ability to help students understand and create CAs. We plan to adapt our materials to more formal educational settings and develop more curriculum on constrained vs. unconstrained NL. Overall, we hope that CONVO can bring us one step closer to empowering anyone to build CAs that can solve the problems of tomorrow. CONVO's source code can be found on GitHub [25].

ACKNOWLEDGMENTS

We would like to thank the entire MIT App Inventor team, Hal Abelson and Jeffrey Schiller in particular. We would also like to thank all of the students who participated in our workshop for their time and ideas.

REFERENCES

[1] The Next Web, "Bizarre video of Alexa (not) talking about the cia sparks conspiracy theories." https://thenextweb.com/news/amazon-alexa-cia-conspiracy/amp. Accessed: 2021-04-23.

[2] Tip Hero, "19 of the creepiest things Alexa has ever said or done." https://tiphero.com/creepiest-things-alexa-has-done. Accessed: 2021-04-23.

[3] Reddit, "Is Alexa spying on us? we're too busy to care — and we might regret that." https://www.reddit.com/r/technology/comments/6t0x0k/is_alexa_spying_on_us_were_too_busy_to_care_and/. Accessed: 2021-04-23.

[4] G. A. Fowler, "I live with Alexa, Google Assistant and Siri. here's which one you should pick.." https://www.washingtonpost.com/technology/2018/11/21/i-live-with-alexa-google-assistant-siri-heres-which-you-should-pick/. Accessed: 2021-04-06.

[5] K. Kozuch, "Alexa vs. Google Assistant vs. Siri: Which smart assistant is best?." https://www.tomsguide.com/us/alexa-vs-siri-vs-google,review-4772.html. Accessed: 2021-04-06.

[6] D. Long and B. Magerko, "What is ai literacy? competencies and design considerations," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, (New York, NY, USA), p. 1–16, Association for Computing Machinery, 2020.

[7] D. Touretzky, C. Gardner-McCune, F. Martin, and D. Seehorn, "Envisioning ai for k-12: What should every child know about ai?," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 9795–9799, Jul. 2019.

[8] J. Van Brummelen, T. Heng, and V. Tabunshchyk, "Teaching tech to talk: K-12 conversational artificial intelligence literacy curriculum and development tools," in *2021 AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI)*, (Online), AAAI, 2021.

[9] E. K. Morales-Urrutia, J. M. Ocaña, and D. Pérez-Marín, "How to integrate emotions in dialogues with pedagogic conversational agents to teach programming to children," *Innovative Perspectives on Interactive Communication Systems and Technologies*, p. 66, 2020.

[10] N. A. Mack, D. G. M. Rembert, R. Cummings, and J. E. Gilbert, "Co-designing an intelligent conversational history tutor with children," in *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, IDC '19, (New York, NY, USA), p. 482–487, Association for Computing Machinery, 2019.

[11] T. Komatsubara, M. Shiomi, T. Kanda, H. Ishiguro, and N. Hagita, "Can a social robot help children's understanding of science in classrooms?," in *Proceedings of the Second International Conference on Human-Agent Interaction*, HAI '14, (New York, NY, USA), p. 83–90, Association for Computing Machinery, 2014.

[12] G. Biswas, J. R. Segedy, and K. Bunchongchit, "From design to implementation to practice a learning by teaching system: Betty's brain," *International Journal of Artificial Intelligence in Education*, vol. 26, no. 1, pp. 350–364, 2016.

[13] P. Lin, J. Van Brummelen, G. Lukin, R. Williams, and C. Breazeal, "Zhorai: Designing a conversational agent for children to explore machine learning concepts.," in *AAAI*, pp. 13381–13388, 2020.

[14] D. Duran, "Learning-by-teaching. evidence and implications as a pedagogical mechanism," *Innovations in Education and Teaching International*, vol. 54, no. 5, pp. 476–484, 2017.

[15] J. Van Brummelen, K. Weng, P. Lin, and C. Yeo, "Convo: What does conversational programming need?," in *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pp. 1–5, 2020.

[16] D. Rough and B. Cowan, "Poster: Apis for ipas? towards end-user tailoring of intelligent personal assistants," in *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pp. 1–2, 2020.

[17] Amazon, "Alexa Skills Kit." https://developer.amazon.com/en-US/alexa/alexa-skills-kit, 2020. Accessed: 2021-04-24.

[18] Google, "Actions on google." https://developers.google.com/assistant, 2020. Accessed: 2021-04-24.

[19] Amazon, "Alexa skill blueprints." https://blueprints.amazon.com/, 2020. Accessed: 2021-04-24.

[20] J. Van Brummelen, "Tools to create and democratize conversational artificial intelligence," Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, 2019.

[21] M. G. Helander, *Handbook of human-computer interaction*. Elsevier, 2014.

[22] J. Good and K. Howland, "Programming language, natural language? supporting the diverse computational activities of novice programmers," *Journal of Visual Languages & Computing*, vol. 39, pp. 78–92, 2017.

[23] T. J.-J. Li, T. Mitchell, and B. Myers, "Interactive task learning from GUI-grounded natural language instructions and demonstrations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, (Online), pp. 215–223, Association for Computational Linguistics, July 2020.

[24] J. Van Brummelen, "Conversational agents to democratize artificial intelligence," in *2019 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pp. 239–240, 2019.

[25] J. Zhu, K. Weng, and J. Van Brummelen, "CONVO." https://github.com/jessvb/convo, 2019.

[26] J. Van Brummelen, V. Tabunshchyk, and T. Heng, ""Alexa, can i program you?": Student perceptions of conversational artificial intelligence before and after programming Alexa," in *Proceedings of the 2017 Conference on Interaction Design and Children*, IDC '21, (New York, NY, USA), p. to appear, Association for Computing Machinery, 2021.

[27] S. Schöbel, A. Janson, and A. Mishra, "A configurational view on avatar design–the role of emotional attachment, satisfaction, and cognitive load in digital learning," in *Fortieth International Conference on Information Systems, Munich*, 2019.

[28] J. B. Wiggins, J. F. Grafsgaard, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "Do you think you can? the influence of student self-efficacy on the effectiveness of tutorial dialogue for computer science," *International Journal of Artificial Intelligence in Education*, vol. 27, no. 1, pp. 130–153, 2017.

[29] "Rasa," 2020. https://rasa.com/, Last accessed on 2020-09-26.

[30] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[31] K. Kvale, O. A. Sell, S. Hodnebrog, and A. Følstad, "Improving conversations: Lessons learnt from manual analysis of chatbot dialogues," in *Chatbot Research and Design* (A. Følstad, T. Araujo, S. Papadopoulos, E. L.-C. Law, O.-C. Granmo, E. Luger, and P. B. Brandtzaeg, eds.), (Cham), pp. 187–200, Springer International Publishing, 2020.

[32] "Spark!." https://esp.mit.edu/teach/Spark/index.html. Accessed: 2021-05-31.

[33] J. Zhu and J. Van Brummelen, "CONVO Appendices 2021." https://gist.github.com/jessvb/7c091e9e8f22e1125b56b1da9d495dc5, 2021.

[34] Zygomatic, "Wordclouds.com." https://www.wordclouds.com/. Accessed: 2021-05-06.

[35] G. Sullivan and A. Artino, "Analyzing and interpreting data from likert-type scales," *Journal of graduate medical education*, vol. 5, pp. 541–2, 12 2013.