Integrating Speech Technology into the iSTART-Early Intelligent Tutoring System

Renu Balyan[1] (✉), Tracy Arner[2], Tong Li[2], Ellen Orcutt[3],
Reese Butterfuss[2], Panayiota Kendeou[3], and Danielle McNamara[2]

[1] State University of New York at Old Westbury, Old Westbury, NY, USA
balyanr@oldwestbury.edu
[2] Arizona State University, Arizona, USA
[3] University of Minnesota, Minneapolis, MN, USA

Abstract. Speech technology (automated speech recognition – ASR and text-to-speech) offers great promise in the field of automated literacy and reading tutors for children. Students in third and fourth grades struggle with generating longer strings of text on a QWERTY keyboard because they still “hunt and peck” for the letters and symbols rather than typing fluently. Thus, in addition to reading comprehension, students’ performance is limited by their ability to translate their ideas into language and then transcribe those words into written text. Fourth grade students produce fewer words and recall less when typing or writing a response relative to speaking. Hence, speech technology is a crucial component in the development of iSTART-Early, an intelligent tutoring system that aims to provide online, automated reading strategy instruction and practice to improve deep comprehension for third and fourth graders. This paper discusses the key components and features of the speech technology incorporated into iSTART-Early. We also discuss some initial findings from pilot studies conducted with adults and youth for this ASR integrated tutoring system. Finally, we discuss considerations for future development of speech technology and integration with intelligent tutoring systems.

Keywords: Automated speech recognition · Text-to-speech · Intelligent tutoring systems

1 Introduction

The iSTART-Early project aims to develop an online, automated reading strategy tutor that provides instruction and practice designed to improve deep comprehension for students in 3rd and 4th grades. iSTART-Early builds upon a previously designed tutoring system (iSTART). iSTART was developed for relatively proficient readers, namely high school students and readers who were already able to read and type in their responses (i.e., self-explanations, summaries, and questions) using a standard QWERTY keyboard. iSTART is not appropriate for elementary school students in its current form because

S. Crossley and E. Popescu (Eds.): ITS 2022, LNCS 13284, pp. 1–9, 2022.
https://doi.org/10.1007/978-3-031-09680-8_34

students in 3rd and 4th grades may not be proficient readers nor are they proficient typists on a standard keyboard. Therefore, we are developing a new system that would offer students the *option to be able to read aloud the text* or *only the words they are struggling to identify* [1] by a pedagogical agent, as well as the *ability to speak aloud their explanations* and *edit their answers* generated by the automatic speech recognition system. Providing such scaffolding is imperative because 3rd and 4th grade students need options to read the text aloud and speak their responses while learning reading comprehension strategies. Students in 3rd and 4th grades struggle with generating longer strings of text on a keyboard because they still "hunt and peck" for the letters and symbols that they need rather than typing fluently [2]. Thus, in addition to reading comprehension, students' performance is limited by their ability to *translate* their ideas into language and then *transcribe* those words into written text [3]. Fourth grade students produce fewer words and recall less when typing or writing a response relative to speaking [4]. Hence, the incorporation of speech technology into iSTART-Early constitutes a crucial feature to support students' learning.

1.1 iSTART-Early

iSTART-Early is an intelligent tutoring system (ITS) that provides automated instruction and practice on higher-order reading comprehension strategies to 3rd and 4th grade students. iSTART-Early provides personalized, interactive, game-based strategy instruction on and practice with reading comprehension strategies with grade-appropriate informational texts. iSTART-Early provides students with real-time feedback and instruction on improving the use of effective comprehension strategies. Thus, each student receives personalized instruction from the system. Natural language processing (NLP) combined with speech technologies (automated speech recognition and text-to-speech) enable student interactions with immediate feedback.

The key components of iSTART-Early are encapsulated in a meta-game set in space in which students are space travelers traversing five planets based on target reading strategies modified from SERT [5]. The reading strategies are organized on planets named for each one: Ask it (question asking), Reword it (paraphrasing), Find it (main idea identification), Explain it (self-explanation), and Summarize it (summarization). Each planet includes: a) video lessons providing guided instruction and demonstration of the five targeted reading strategies; b) practice games to increase engagement when interacting with the system and provide opportunities for deliberate practice; c) automated speech recognition (ASR) capabilities that allow students to practice reading strategies without the additional burden of typing; d) text-to-speech functionality that assists students' reading by having the texts or difficult words read aloud to them; and e) a teacher interface that allows teachers to assign texts and monitor students' performance to provide additional support and feedback when necessary, creating blended-learning opportunities. This paper focuses on describing the speech technologies (ASR and text-to-speech) and the complete workflow of these components.

2 Integration of Speech Technologies into iSTART-Early

The speech technologies were integrated into iSTART-Early to provide several functionalities that allow students to: a) have difficult (predefined) words or instructions read aloud by a pedagogical agent in English, Spanish, or Chinese, using text-to-speech; b) speak aloud their responses (e.g., paraphrases, questions, self-explanations) using ASR instead of typing responses; and c) edit and update the transcription generated by the ASR engine. The system also includes the capability to filter out and clean frozen expressions, curse words, and implements a spell checker.

2.1 The Process Flow

The overall process flow for the ASR and text-to-speech modules integrated within iSTART-Early is as follows (also see Fig. 1):

1) The student is shown the text to be self-explained, summarized etc. after logging into iSTART-Early. The student has an option to let the pedagogical agent read aloud the entire text or some predefined difficult words and their definitions. This feature is implemented via the text-to-speech component.
2) The student can also choose to type or speak aloud their response after logging in. The web API asks for permission to use the microphone, if the student chooses to speak their response.
3) The student is prompted to speak aloud the response, once the microphone permissions are accepted and asked to click the stop button once done speaking to stop the recording.
4) The student is presented with the ASR transcribed text and given the option to edit the ASR generated output. The student then submits the response.
5) The response generated by the ASR system and edited by the student is automatically filtered of spelling errors, frozen expressions, and curse words prior to algorithmic evaluation and the generation of pedagogical feedback.

2.2 Automated Speech Recognition Interface

Google Web Speech (GWS) API was used to create the ASR functionality. GWS provides cloud-based voice transcription services that transcribes speech into text in real-time. GWS relies on google chrome web browser to communicate and transfer data between the user interface and the GWS server. The acoustic models (statistical representation of sounds that make up words) for GWS were trained and developed using 5000 h of data that showed very high recognition rate for people with speech disorder [6]. The GWS in the iSTART-Early was implemented by adding a JavaScript event listener to the microphone button within the interface to process ASR requests. Once the connection request is detected, the interface creates a GWS controller to monitor user speech and transfer the audio data. The controller is responsive to “pause” or “stop” requests by the user.

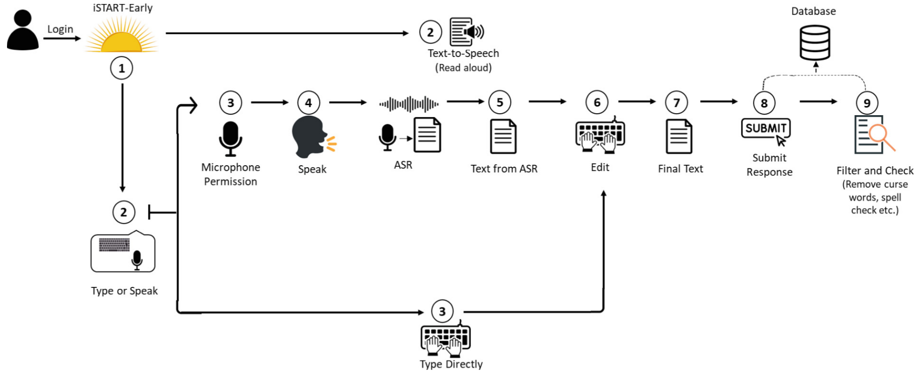
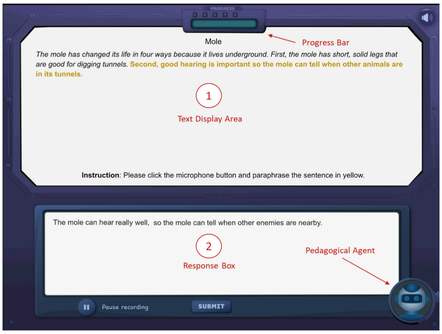


Fig. 1. The speech modules (ASR and text-to-speech) process flow

The student interface includes ASR to allow students to practice reading strategies without the additional burden of typing [7]. The ASR functionality added to the paraphrasing practice module is illustrated in Fig. 2. The interface consists of several components: a) a *text display* area (top section of the interface, marked as (1) - displays the text and target sentences that the students are tasked for paraphrasing; b) a *progress bar* (above the text display area) - indicates students' task completion progress; c) a *response box* (input box below the text display area, marked as (2) - shows the response typed by the student or the ASR transcription of the student's spoken response; d) a *pedagogical agent* (robot at the bottom-right corner) - provides students with instructions (spoken and text-based) for the tasks and step-by-step guidance on how to use the ASR; e) a *microphone button* (bottom-left of the response box) – is clicked to activate the ASR, which changes to a pause button once activated; and f) a *submit button* (bottom-middle of the response box) - to submit the final response.

2.3 Text-To-Speech: Audio Collection Integration

The text-to-speech component has been added to the student interface to request definitions for pre-identified vocabulary words or request that the entire text is read aloud by the pedagogical agent. We have utilized the pre-existing iSTART system module and created a set of new databases for all the texts and predefined words definitions. The E-learning authoring tool (Articulate Storyline) was used to generate MP3 audio files of texts and definitions for the pre-identified and difficult words. Figure 3 illustrates the text-to-speech functionality in a game-based *question-asking* strategy practice activity. The system retrieves the audio for the text from the database and plays it when the user clicks the speaker button in the middle of the interface (highlighted in blue). The users have the options to pause and replay the audio. The vocabulary text-to-speech functionality embedded in the interface (underlined word highlighted in red) is shown in Fig. 3. The users can hover over the underlined word to learn its pronunciation and definition. These definitions can also be read aloud by using the text-to-speech functionality.



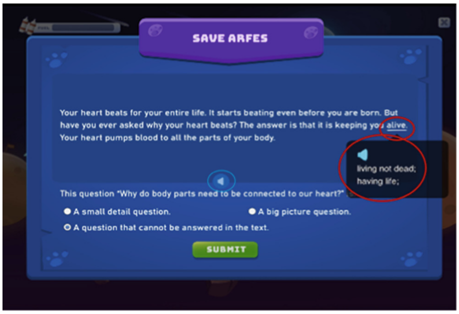
Mole

The mole has changed its life in four ways because it lives underground. First, the mole has short, solid legs that are good for digging tunnels. **Second, good hearing is important so the mole can tell when other animals are in its tunnels.**

Instruction: Please click the microphone button and paraphrase the sentence in yellow.

The mole can hear really well, so the mole can tell when other enemies are nearby.

Fig. 2. Paraphrase module interface with ASR embedded technology



Your heart beats for your entire life. It starts beating even before you are born. But have you ever asked why your heart beats? The answer is that it is keeping you alive. Your heart pumps blood to all the parts of your body.

living not dead; having life;

This question "Why do body parts need to be connected to our heart?"

○ A small detail question. ○ A big picture question.

◉ A question that cannot be answered in the text.

Fig. 3. Text-to-speech and difficult word vocabulary functionality

2.4 Key Features of the ASR and Text-To-Speech Interface

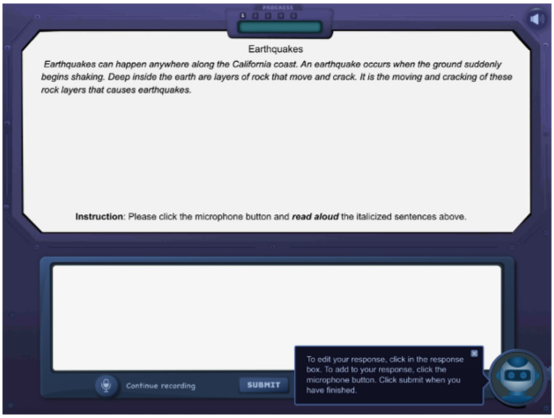
User Instruction. To help users become acquainted with the speech components of the system and provide smooth experience, several guiding components for the system were designed, including an *On-boarding Walk-Through*, *UI Status Indicator*, and *Action Reminder*.

On-Boarding Walk-Through. In the "walk-through" tutorial at the beginning of the system for first-time users, the pedagogical agent (a friendly robot named "Robo") introduces features to the users after they log into the system. Following the feature introduction, Robo demonstrates the steps to activate the ASR and how it can be used to generate responses. Robo also instructs users to explicitly verbalize punctuation for each sentence, because the current ASR technology is not capable of automatically adding punctuation while transcribing. Robo then informs the user to identify any ASR-introduced transcription errors that may need correction. Finally, Robo demonstrates how to pause the ASR and edit responses.

UI Status Indicator. Once users permit the ASR system to use their microphone, a red circular marker (dot) appears on the web page tab that notifies users that they are being recorded. To further clarify the recording status, audio reminders such as "recording started" and "recording paused" were added to specify when the user changes recording status. In addition, when users hover over any button, a dialog box appears and displays information explaining the associated function.

Action Reminder. The pedagogical agent offers feedback to remind the user of the next steps, when using the ASR to produce responses. For instance, after the user switches from the recording mode to the editing mode, the robot instructs the user on the possible actions that can be taken (see Fig. 4 bottom-right corner).

Editing Features. The editing feature enables users to pause ASR and edit the transcriptions generated from their spoken responses. To switch from recording mode to editing mode, users can click the microphone button or in the response box. While in editing mode, users can edit the transcribed text or add text using their keyboard or by reactivating their microphone and re-engaging the ASR technology. Figure 4. Action reminders by "Robo".



Earthquakes

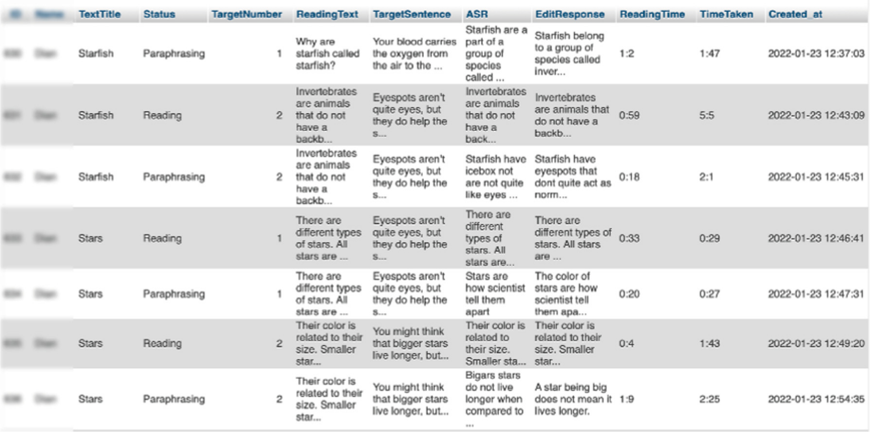
Earthquakes can happen anywhere along the California coast. An earthquake occurs when the ground suddenly begins shaking. Deep inside the earth are layers of rock that move and crack. It is the moving and cracking of these rock layers that causes earthquakes.

Instruction: Please click the microphone button and ***read aloud*** the italicized sentences above.

To edit your response, click in the response box. To add to your response, click the microphone button. Click submit when you have finished.

Fig. 4. Action reminders by "Robo"

Database Structure. The database was designed to collect information related to the user interactions with the system, particularly the speech technology needed to evaluate user responses generated by the ASR. Five different types of information are collected and stored in the database (Fig. 5): a) *text information* - includes the prompt text displayed to users (i.e., text title, text script, and target sentences); b) *original ASR transcripts* – all transcripts generated by the ASR, including ASR transcribed errors; c) *user-edited transcripts* - the edited transcripts for each practice session are stored; d) *user voice recordings* - user speech is recorded and stored as audio files in the database while they speak aloud the response; and e) *time-related features* – time taken by the user to read the text and edit the responses is also recorded and stored.



[redacted]	[redacted]	TextTitle	Status	TargetNumber	ReadingText	TargetSentence	ASR	EditResponse	ReadingTime	TimeTaken	Created_at
[redacted]	[redacted]	Starfish	Paraphrasing	1	Why are starfish called starfish?	Your blood carries the oxygen from the air to the ...	Starfish are a part of a group of species called ...	Starfish belong to a group of species called inver...	1:2	1:47	2022-01-23 12:37:03
[redacted]	[redacted]	Starfish	Reading	2	Invertebrates are animals that do not have a backb...	Eyespots aren't quite eyes, but they do help the s...	Invertebrates are animals that do not have a back...	Invertebrates are animals that do not have a backb...	0:59	5:5	2022-01-23 12:43:09
[redacted]	[redacted]	Starfish	Paraphrasing	2	Invertebrates are animals that do not have a backb...	Eyespots aren't quite eyes, but they do help the s...	Starfish have icebox not are not quite like eyes ...	Starfish have eyespots that dont quite act as norm...	0:18	2:1	2022-01-23 12:45:31
[redacted]	[redacted]	Stars	Reading	1	There are different types of stars. All stars are ...	Eyespots aren't quite eyes, but they do help the s...	There are different types of stars. All stars are...	There are different types of stars. All stars are ...	0:33	0:29	2022-01-23 12:46:41
[redacted]	[redacted]	Stars	Paraphrasing	1	There are different types of stars. All stars are ...	Eyespots aren't quite eyes, but they do help the s...	Stars are how scientist tell them apart	The color of stars are how scientist tell them apa...	0:20	0:27	2022-01-23 12:47:31
[redacted]	[redacted]	Stars	Reading	2	Their color is related to their size. Smaller star...	You might think that bigger stars live longer, but...	Their color is related to their size. Smaller sta...	Their color is related to their size. Smaller star...	0:4	1:43	2022-01-23 12:49:20
[redacted]	[redacted]	Stars	Paraphrasing	2	Their color is related to their size. Smaller star...	You might think that bigger stars live longer, but...	Bigars stars do not live longer when compared to ...	A star being big does not mean it lives longer.	1:9	2:25	2022-01-23 12:54:35

Fig. 5. Screenshot of the user interaction data stored in the database

3 Initial Evaluation of Speech Technologies into ISTART-Early

3.1 Adult Pilot Study

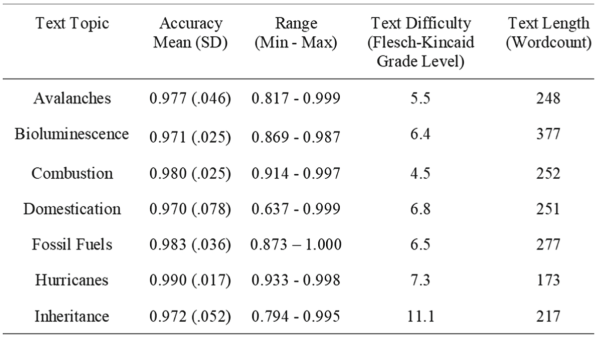
Participants and Procedure. Participants accessed the study via Amazon's MTurk service. The original sample consisted of 38 adults; however, 5 participants were excluded from the dataset due to incomplete or irrelevant responses or poor response quality. Therefore, the final sample consisted of 33 participants (17 male, 16 female). The mean age was 39 years (Standard Deviation, $SD = 10$); 23 participants were White, 5 Black, 1 Asian, 1 Hispanic, and 2 were multiracial; 19 participants reported earning a four-year college degree, 6 a graduate degree, 5 had a high school diploma, and 4 reported attending some college. All participants reported that they were fluent in English. Participants first completed the ASR task that instructed participants to read the texts aloud at a comfortable pace while the ASR system transcribed their utterances. Participants were not permitted to edit the transcriptions. Upon completion, they were directed to Qualtrics [8] to complete the demographics questionnaire.

Texts. The current study included seven expository texts selected from a larger body of texts commissioned for the development of the iSTART-Early system. The texts used in this study were chosen based on word selection (i.e., number and type of content words), syntax, and punctuation. Texts were presented, one at a time, as individual blocks of text without paragraph breaks. See Table 1 for text length and Flesch-Kincaid Grade Level [9] scores.

Measures. Accuracy of the ASR system was assessed using the Word2Vec source overlap using the Tool for the Automatic Analysis of Text Cohesion (TAACO) [10].

Analysis. We examined the overlap between the participants' ASR responses in order to evaluate the accuracy of the ASR system. The Word2Vec overlap provides an indicator of the accuracy of the ASR system, such that higher overlap values indicate more accurate transcriptions. Word2Vec [11] relies on a neural-network model to represent words

Table 1. ASR accuracy and text characteristics



Text Topic	Accuracy Mean (SD)	Range (Min - Max)	Text Difficulty (Flesch-Kincaid Grade Level)	Text Length (Wordcount)
Avalanches	0.977 (.046)	0.817 - 0.999	5.5	248
Bioluminescence	0.971 (.025)	0.869 - 0.987	6.4	377
Combustion	0.980 (.025)	0.914 - 0.997	4.5	252
Domestication	0.970 (.078)	0.637 - 0.999	6.8	251
Fossil Fuels	0.983 (.036)	0.873 – 1.000	6.5	277
Hurricanes	0.990 (.017)	0.933 - 0.998	7.3	173
Inheritance	0.972 (.052)	0.794 - 0.995	11.1	217

and phrases in a vector-space model. Words co-occurring in similar contexts are represented as being closer, whereas words with dissimilar contexts are represented as being farther apart, in different regions of the vector space. Overlap scores reflected the cosine similarity between the vectors corresponding to the compared segments (e.g., sentences, paragraphs, or documents), which were obtained by summing the vector weights for each word [12].

Results. With respect to overall accuracy, descriptive analyses revealed an overall mean Word2Vec overlap score, across all texts of .979 (*SD* = .044). Accuracy scores were generally high across all seven texts (see Table 1). Thus, results suggest that the ASR system accurately transcribed adult readers' speech as they read expository texts. However, it is critical to examine the ASR system's accuracy for young students because they are the target users for this new ITS i.e., iSTART-Early.

In addition to the adult pilot study, two late elementary students were recruited from a summer tutoring group in the Midwest with parental consent to test various features of the speech modules. Students generally reported positive experiences using the ASR system. The students were found to be fairly fluent readers. The results of these analyses also suggested that the ASR system transcribed students' reading fairly accurately. When errors were qualitatively analyzed some notable sources of error were found to stem from *compound words, homophones, and punctuations such as quotes and parentheses.*

4 Future Work/Next Steps

Our initial test of the ASR system was promising. The ASR transcription accuracy for fluent, adult readers suggests that errors occurring with students could be a result of their less fluent reading skills. However, it is important to note that the youth ASR pilot was conducted with only two students. Therefore, it is necessary to conduct a larger study with children to gain a better understanding of the biggest contributors of low accuracy. In addition to gathering a larger sample of read aloud data, we also need to evaluate the ASR system accuracy when students are generating responses such as verbalizing a self-explanation, a question, or a summary. The generation of constructed responses differs from read aloud activities because the students will be turning the microphone on and off as they generate thoughts and will have the opportunity to edit incorrect words that they speak or words that were transcribed incorrectly. Future studies will investigate the accuracy of students' read aloud protocols as well as generated responses in an effort to ensure that the ASR system is both accurate and sufficiently easy for students to use successfully. Further, we will collect keystroke and log data to better understand what types of errors students are making, what kinds of errors they choose to correct, and when they do so (i.e., as soon as the error is made or when they are done). The more we understand the error and correction patterns that frequently occur, the better we will be able to adapt the system to meet the needs of young learners.

Acknowledgements. The authors would like to recognize the support of the Institute of Education Sciences (IES), U.S. Department of Education through Grant R305A190050, the National Science Foundation - NSF through Award# 2131052, and the Office of Naval Research - ONR through Grant N000142012623. The opinions expressed are those of the authors and do not represent views of the IES, the NSF, or the ONR.

References

1. Mayer, R.E.: Multimedia Learning, 2nd edn. Cambridge University Press, New York (2009)
2. Wijekumar, K.K., Meyer, B.J., Lei, P.: Large-scale randomized controlled trial with 4th graders using intelligent tutoring of the structure strategy to improve nonfiction reading comprehension. Educ. Technol. Re. Dev. **60**(6), 987–1013 (2012)
3. Berninger, V.W., Abbott, R.D., Augsburger, A., Garcia, N.: Comparison of pen and keyboard transcription modes in children with and without learning disabilities. Learn. Disabil. Q. **32**, 123–142 (2009)
4. Bourdin, B., Fayol, M.: Is written language production more difficult than oral language production: a working-memory approach. Int. J. Psychol. **29**, 591–620 (1994)
5. McNamara, D.S.: SERT: self-explanation reading training. Discourse Process. **38**(1), 1–30 (2004)
6. Anggraini, N., Kurniawan, A., Wardhani, L.K., Hakiem, N.: Speech recognition application for the speech impaired using the android-based google cloud speech API. Telkomnika **16**(6), 2733–2739 (2018)
7. Hagen, A., Pellom, B., Cole, R.: Highly accurate children's speech recognition for interactive reading tutors using subword units. Speech Commun. **49**(12), 861–873 (2007)
8. Qualtrics: Version (2020). Qualtrics (2005). https://www.qualtrics.com
9. Flesch, R.: Flesch-Kincaid readability test, **26**(3) (2007). Accessed October 2007
10. Crossley, S.A., Kyle, K., McNamara, D.S.: The tool for the automatic analysis of text cohesion (TAACO): automatic assessment of local, global, and text cohesion. Behav. Res. Methods **48**(4), 1227–1237 (2016)
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, pp. 3111–3119. Curran Associates, Red Hook (2013)
12. Crossley, S.A., Kyle, K., Dascalu, M.: The tool for the automatic analysis of cohesion 2.0: integrating semantic similarity and text overlap. Behav. Res. Methods **51**(1), 14–27 (2018). https://doi.org/10.3758/s13428-018-1142-4