15 August 18

Evaluation of Enskill® English at the University of Novi Sad, Serbia

W. Lewis Johnson, *Alelo Inc.*

6171 W. Century Blvd., Suite 360, Los Angeles, CA 90045 USA.

ljohnson@alelo.com

INTRODUCTION

Of the four main language skills—speaking, listening, reading, and writing—speaking is by far the most important according to a recent survey of English teachers worldwide (Thiriau, 2017). Yet English language learners (ELLs) around the world struggle to develop their speaking and listening skills. They have limited opportunity to practice with native English speakers; instead, they must practice with their English teacher or with other students who themselves have limited spoken proficiency for the same reasons. The problem of spoken proficiency is particularly acute for students in large classes where speaking practice activities are hard to organize and manage.

Alelo Enskill® English gives learners opportunities to practice their spoken English skills in realistic conversations with a variety of simulated native speakers in the form of animated characters that speak and understand English. They can practice as much as they want in a safe environment without fear of making embarrassing mistakes. This helps build proficiency and self-confidence. It also reduces anxiety about speaking, which has repeatedly been shown to explain a significant fraction of the variance in foreign language achievement as well as student attrition (Bailey et al., 2003). Hsieh (2008) studied differences between successful and unsuccessful college-level students of European languages, and found attitude toward foreign language and anxiety about speaking in class to be strong predictors of poor academic performance.

A snapshot evaluation tested Enskill English with students in an English for specific purposes (ESP) program at the University of Novi Sad in Serbia. We tested the effectiveness of the released version of Enskill English, tested the performance of a new version of the Enskill dialogue system that had yet to be released, and analyzed samples of learner speech to inform future extensions to Enskill. The evaluation provided evidence that Enskill is helpful for learning spoken English skills; detailed analysis of learner performance indicated that learner performance improved through repeated practice. It indicated that Enskill can be extended to meet the needs of ESP students who want to learn to speak English in a professional context.

OVERVIEW OF THE STUDY

The study took place in April and May of 2018. The study population was a class of intermediate-level English students in an English for special purposes (ESP) program at the Faculty of Technical Sciences of the University of Novi Sad in Serbia. One class of eighty (80) students participated in the study. One student was a native of Bosnia and Herzegovina, the rest were natives of Serbia. All students had previously taken and passed a CEFR B1 English course. However, according to their instructor the students were in need of additional spoken English practice. Due to the large class size these students had few opportunities to practice speaking English in class.

At the time of the evaluation the CEFR (Common European Framework of Reference) A1 level of Enskill English had already been used extensively by beginner students, and the A2 level was nearing completion. The purpose of the evaluation was as follows:

- To evaluate whether Enskill English is beneficial for intermediate-level ELLs.
- To test a new version of the Enskill dialogue system and compare it to the released version.
- To collect data from intermediate-level ELLs to inform the development of future extensions of Enskill English for intermediate-level ELLs.

At the intermediate level of spoken English (B1 and B2 on the CEFR scale) learners begin to have the ability to use English on the job and in professional settings. Most ESP programs in industries such as information technology, healthcare, hospitality, law enforcement, and aviation target intermediate-level learners. Thus, for many learners intermediate-level spoken proficiency is critical for job advancement and professional growth.

The director of the English program at the University of Novi Sad, Ms. Vesna Bulatović, was willing to cooperate with Alelo on a pilot evaluation and collect student surveys to supplement the user interaction data collected by Enskill. This allowed the comparison of learner judgements and reactions against findings from analyzing the user interaction learner data. The students in the Novi Sad program have limited opportunities to practice speaking English, so we hypothesized that the existing A1 and A2 level simulations might be beneficial for them. We wished to test the following hypotheses:

- **Hypothesis 1.** ELLs at the CEFR B level consider Enskill English to be a good way to practice English. They find it useful, fun, and easy to use.
- **Hypothesis 2.** ELLs at the CEFR B level benefit from practice with Enskill English A-level simulations.

Since previous evaluations of Enskill English with beginner-level English students at Laureate International Universities generated positive responses, we expected this study to support Hypothesis 1. The status of Hypothesis 2 was much less certain, however. The A-level simulations were not designed for B-level learners, and we were not sure how B-level learners would react. The following are possible reasons why Hypothesis 2 might be true:

- Students can use simulations for review, to maintain mastery of language skills covered earlier in the language course.
- They can practice vocabulary and structures common to real-world situations.

The following are some possible reasons why the study might not confirm Hypothesis 2:

- B-level learners might regard the A-level conversations to be too easy for them.
- B-level learners might say complex utterances or discuss topics that the A-level simulations were not designed to handle.

The study had an additional objective: to test a new version of Enskill English's natural-language processing (NLP) pipeline. The new NLP pipeline incorporates statistical natural-language processing technology that enables it to understand the intent behind a wider range of learner utterances. We expected that it would be particularly useful for intermediate-level learners who have a more flexible speaking ability than beginner ELLs. We therefore decided to run the new NLP pipeline in the background during the test, and compare its performance against the released version.

The study provided an opportunity to collect a corpus of speech of intermediate-level learners. These data would inform the design of future enhancements of Enskill to support intermediate-level language, and could be used to train further dialogue models. The corpus consisted of the speech of Serbian speakers, and so can be used to retrain Enskill English's models to recognize errors common to Serbian speakers of English.

STUDY MATERIALS

The study materials included two simulation modules at the A1 level and two simulation modules at the A2 level. The A1 simulations had been previously released and had already been used extensively by beginner-level students. The A2 simulations were still beta versions that were in the process of final testing.

During the trial, two versions of Enskill's natural language dialogue system ran simultaneously. The foreground system, which drove the simulated native speakers' responses, used Alelo's partial-matching algorithm to interpret learner utterances by matching them against a library of possible utterances and a library of common learner errors. The error library included errors from speakers of several languages (Portuguese, Spanish, Thai, and Turkish), but none from Serbian speakers. The background system also included statistical natural language processing.

Enskill kept track of the number of times each student attempted each simulation, and whether the student successfully completed all the task objectives. It also recorded the date and time of each student's first and last attempt, as well as the total time spent. We provided these statistics to the instructor so she could track the students' usage of the courseware.

The pilot lasted three weeks. During the study period the students were directed to practice each simulation until they could complete all the objectives. They were free to practice the simulations more times if they wished.

After the students completed the pilot they completed a survey of their attitudes toward Enskill English and the simulations. The survey included 5-point Likert questions asking if the exercises were a good way to practice English, were engaging and easy use, and helped them with their English speaking and listening skills. There were also free-form questions in which students could describe what they liked about Enskill English and what they would like to see improved. The survey included a net promoter score question to determine whether they would recommend Enskill English to their family and friends.

EVALUATION RESULTS

A total of 72 students completed the completed the survey. 71 attempted the A1 simulations and 66 attempted the A2 simulations.

The net promoter score (NPS) from the survey was **23**. There were a total of 27 promoters, 28 passives, and 17 detractors. An NPS above zero is considered good, and an NPS of 50 is considered excellent (Keck, 2017), so this result is very good.

The students believed that Enskill English was a good way to practice speaking English (mean=4.03 of a possible 5, s.d.=0.69). Only one student disagreed, and no student strongly disagreed. Most students felt that the interactive conversations were amazing (mean=3.83, s.d.=0.91). Nine students disagreed, but no student strongly disagreed. Most students felt that Enskill English is easy to use (mean=4.06, s.d.=0.85). Two students disagreed and two students strongly disagreed.

Together these findings appear to confirm Hypothesis 1: ELLs in this study considered Enskill English to be a good way to practice English. Further statistical analyses are currently being performed to confirm the statistical significance of these findings.

Most students felt that the interactive conversations helped their English speaking and listening skills (mean=3.49, s.d.=1.09). 8 students disagreed and 5 students strongly disagreed. Thus Hypothesis 2 was confirmed for most students, although there was a small minority of students (18%) who disagreed that the simulations were helpful.

The students who disagreed that Enskill helped them did so mainly because they perceived the dialogue choices to be too limited and restrictive according to their comments and other ratings of the product. This is not surprising because the NLP pipeline was designed to understand the relatively simple language of students at the CEFR A1 level. The tests of the new NLP pipeline indicated that it performs much better, as described below, which should address these students' concerns. Some students also encountered bugs in the A2 simulations, which were beta versions and still undergoing testing.

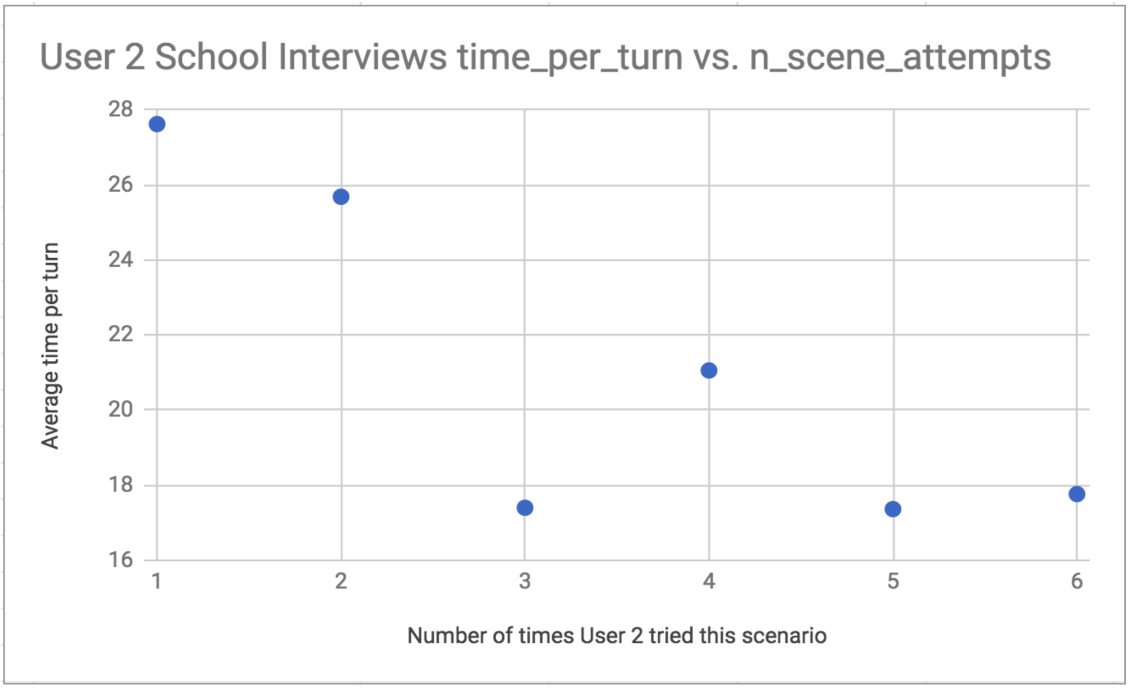

Fig.1. Learner performance increase after repeated practice.

Here are some comments from the students about what they liked about Enskill English:

- It is very imteresting [sic] and it can help you to aprove [sic] your communication skills.
- It is very useful for practice.
- It's very easy to use and you can learn a lot.
- The variety of the conversations.
- There are very interesting conversation [sic] and I really liked it.

Here are some comments from the students about what they would like to see improved:

- Add more variety of answers and possible questions.

- I would add more themes for conversation.
- Find the right activity for you, improve your English writing skills, improve your English reading skills....
- Maybe some more professional conversation.
- I don’t know, everything is fine.

Of the students who participated in the pilot, the total time spent using Enskill English varied greatly (mean=32:31, s.d.=28:15, min=00:24, max=2:41:37). Most students completed each simulation module just once, but many practiced them 3, 4, or as many as 14 times.

We are still analyzing the learner data to measure how student performance improved through repeated practice. Figure 1 shows an example from earlier data that shows how average time per conversational turn decreases with practice.

DATA COLLECTION RESULTS

Analysis of the student speech recordings collected during the study confirmed that the students had an intermediate level of spoken proficiency. Their speech was reasonably fluent but exhibited some disfluencies such as sentence restarts.

Intermediate learners are able to construct sentences on their own in contrast with beginners who rely heavily on memorized phrases. The learners in this study produced a wide variety of utterances, which suggested that they were not relying on memorized phrases. Many utterances had errors in grammar and usage, which also suggests they were constructing utterances on the fly.

The following are transcriptions generated by the speech recognizer of examples of ill-formed utterances. Note that the speech recognizer’s language model is trained on grammatical speech, and so when presented with ungrammatical speech it may make transcription mistakes. In response to the question “Where are you traveling to?”

- Are you going to travel to New York.
- Might want to travel to New York City.

In response to the question “What day would you like to depart?”

- I want to depart at May 1st.
- I would like to leave at first May.
- I have to leave at first May.

Although these learners had grammatical and usage errors, they had relatively few pronunciation errors and the speech recognizer transcribed their speech without much difficulty. Enskill’s existing English speech recognizer and substitution table was adequate for understanding the speech of the Serbian speakers in the subject population. We analyzed the utterance data in depth to find pronunciation errors to include in the substitution table, and found only 64 new substitutions out of a total utterance set of 8637 utterances.

This data provides useful insight about the nature of intermediate-level learner speech, which will inform our future extensions of Enskill English for intermediate-level learners. Meanwhile, the NLP pipeline tests helped assess whether the system is technically capable of processing the speech of intermediate-level learners.

TEST RESULTS

During the study we evaluated two different versions of the natural-language processing pipeline. Version 1 used a substring matcher, a substitution table, and a tolerance algorithm to compare each learner’s utterance against a library of pre-authored utterances and find the most likely match. Version

2 augmented this pipeline with a machine-learned statistical classifier created using Microsoft's Language Understanding Interactive Service (LUIS). The version 2 pipeline had the following advantages over the version 1 pipeline:

- LUIS supports authoring with example utterances. This is intuitive for authors and we expected that it would work well for interpreting learner speech. The substitution table is still used to pinpoint errors and provide feedback.
- It can take into account whole-sentence similarity between the learner's utterances and authored utterances. This lets it recognize the learner's intended meaning in spite of learner errors. This proved to be important because as noted above the intermediate-level learners made many grammatical and usage mistakes.
- Classifier thresholds are trained on labelled data using machine learning. Classification improves as more data is collected and the classifiers are retrained.
- Models are created for each simulation. The context of the situation and of the dialogue makes it easier for the system to understand the learner's intent. For example, in the Train Ticket simulation when the ticket agent asks, "Where are you traveling to?" and the learner responds "Are there trains to New York City?" the model interprets the response as saying that the learner intends to travel to New York City. The ticket agent then responds by saying "I can help you plan your trip to New York", instead of "Yes, there are trains to New York City."

We use *success rate,* the percentage of conversational turns that Enskill can assign meaning to, as a measure of Enskill's ability to understand learner speech. *Clickthrough rate* is the percentage of conversational turns where the learner selected a choice from a menu instead of speaking into the microphone. The failure rate is 100% - success rate - clickthrough rate. Ideally the success rate should be similar to the success rate between a language learner and a native speaker, i.e., high but not 100% since native speakers often have trouble understanding what language learners are trying to say.

Table 1 shows the success rates for the participants in the Novi Sad trial, using the version 1 pipeline. Table 2 shows the success rates for a comparable number of beginner ELLs studying at Laureate Education institutions around the world from May 28, 2018 until August 6, 2018. In these tables Total Turns is the number of conversational turns in the sample. The success rates vary considerably from simulation to simulation, but overall is similar for both groups of students. The success rates generally have room for improvement, although the success rate for the Novi Sad students in the *Class Interview with Lily* simulation was quite high at 84%.

Simulation Name	CEFR Level	Total Turns	Success Rate	Clickthrough Rate
Class Interview with Lily	A1	1001	84%	3%
Helping Owen Plan a Party	A1	1235	63%	6%
Jerry's Spaghetti	A2	375	73%	1%
Train Ticket	A2	1692	54%	17%

Table 1. Success and clickthrough rates for Novi Sad students.

Simulation Name	CEFR Level	Total Turns	Success Rate	Clickthrough Rate
Class Interview with Lily	A1	1174	65%	10%
Helping Owen Plan a Party	A1	1180	66%	10%

Table 2. Success rates and clickthrough rates for Laureate Education students.

Version 2 achieved significantly better success rates, consistently above 80%. It was able to correctly interpret incomplete utterances and word ordering mistakes, including all the examples of ill-formed utterances in the preceding section. It can also understand utterances that are relevant to the conversation but do not follow the expected flow of the dialogue, e.g., answering a question with another question.

Based on these test results, version 2 of the NLP pipeline has since been adopted as the NLP pipeline for the Enskill English A2 simulations, and we are planning further tests to measure the improvements in success rate. We are also planning further tests of version 2 on the A1 simulations. These tests will be part of upcoming snapshot evaluations.

DISCUSSION

Overall the students in this pilot responded positively to Enskill English. Considering that the A-level simulations were designed for learners with A-level spoken proficiency, not B-level learners, the survey results were remarkably positive. The program director at the University of Novi Sad was also very encouraged by the results. Here are some of her comments:

- Enskill simulation is a great education tool that helps students practice speaking in real-life situations.
- I really like the idea of teaching soft-skills through simulations, that would be also useful in teaching languages for specific purposes (Business English for example).

There was a wide variation in usage time during the study. This is to be expected since the learning experience is personalized for each learner. If learners can complete a simulation without difficulty they do not need to spend a long time practicing. Learners who encounter difficulties receive much more content to practice. This helps ensure that all learners use their practice time efficiently, focusing on the skills that they each need to improve.

This was the first test of Enskill English with intermediate-level learners, so not surprisingly it identified some areas for improvement. Many students simply wanted to see a wider range of conversations; the complete Enskill English product provides many more conversations. The students indicated that they wanted the system to understand a wider range of utterances. The new version 2 NLP pipeline addresses this need.

CONCLUSIONS

This study confirmed that intermediate-level ELLs see the benefit of practicing spoken English skills with Enskill English. Most students felt that they benefited from practicing with Enskill English simulations. However, some students felt that the dialogue options were too restrictive; others encountered bugs in the simulations that were still in beta test. We have since corrected the bugs and released the new natural language understanding system that recognizes the intent behind a much wider range of learner utterances. We are in the process of converting more simulations to use this new system. These improvements address the concerns of the minority of students who did not feel that they benefited.

We recommend that instructors of intermediate English students provide Enskill English as an optional resource for practice and review. We particularly recommend the A2-level simulations, which use our new natural language dialogue system and model a variety of real-world situations.

In future work we plan to use the data from intermediate-level learners that we collected, we plan to extend Enskill English to include new intermediate-level language tasks, to help ELLs continue making progress to higher levels of language proficiency. We plan to extend Enskill to provide more analytics to teachers, to make it easier for teachers to integrate Enskill into their classes.

Our findings suggest that Enskill approach has broad application to communication skills. Students and instructors both wanted to see more examples of simulations in business or professional contexts. We see potential for the application of Enskill to teaching soft skills.

ACKNOWLEDGMENTS

The author and the Enskill English development team wish to thank Vesna Bulatović and her students at the University of Novi Sad for participating in this study. We also wish to thank Laureate Education for their permission to use their English simulation content in this study.

REFERENCES

Bailey, P., Onwuegbuzie, A.J., & Daley, C.E. (2003). Foreign language anxiety and student attrition. *Academic Exchange Quarterly*, Summer 2003.

Hsieh, P.H. (2008). Why are college foreign language students' self-efficacy, attitude, and motivation so different? *International Education* 38 (1).

Keck, C. (2017). How to calculate (and understand) your net promoter score. Retrieved Aug. 6, 2018 from https://www.promoter.io/blog/calculate-net-promoter-score/.

Thiriau, C. (2017). Teaching Speaking in ELT. Cambridge: Cambridge University Press. Retrieved July 27, 2018 from http://www.cambridge.org/elt/blog/wp-content/uploads/2017/11/Cambridge-global-teaching-speaking-survey-2017.pdf