Initial Evaluations of Adaptive Training Technology for Language and Culture

W. Lewis Johnson, Brenda Lindsay
Alelo Inc.
Los Angeles, CA
ljohnson@alelo.com
blindsay@alelo.com

Andrew Naber, Alan Carlin, Jared Freeman
Aptima
Woburn, MA
anaber@aptima.com, acarlin@aptima.com
jfreeman@Aptima.com

ABSTRACT

The Department of Defense has 40,000 positions requiring foreign language skills, yet 70% are filled by people who lack the necessary language proficiency. Too many language learners fail to reach the desired level of proficiency, or lose proficiency if they do not have continuing opportunities to practice. We propose that adaptive training technology can address this problem and help learners quickly develop and maintain language proficiency. To do this, adaptive training should (a) provide learners opportunities to practice using the language in realistic situations, (b) identify gaps in the learners' language competencies in the context of language use, , and (c) recommend personalized exercises to overcome these gaps. Such an approach offers significant advances over conventional methods that emphasize passive language skills (reading, listening) and basic knowledge of vocabulary and grammar without realistic practice in using the language.

We have implemented this approach in ALLEARN (Adaptive Language LEARNing), an open architecture for language learning and retention. ALLEARN provides a simulation-based training environment for practicing language use and assessing language competencies, and recommends personalized learning trajectories that focus on overcoming each learner's individual competency gaps. It is designed to be used in a blended learning curriculum, supplementing classroom instruction.

To obtain formative feedback and identify best practices for designing adaptive training, focus groups were conducted with Modern Standard Arabic (MSA) students and instructors at the US Army JFK Special Warfare Center and School. Two classes of MSA students were assigned to either the adaptive version of ALLEARN or a non-adaptive version which followed a fixed curriculum sequence. The study investigated a variety of outcomes between these conditions, such as improved performance on oral and written examinations, time-on-task, and training usage behaviors in the context of blended learning. The findings offer implications for best practices of competency-based adaptive training and language learning technology.

ABOUT THE AUTHORS

Dr. W. Lewis Johnson Dr. W. Lewis Johnson co-founded Alelo in 2005 as a spinout of the University of Southern California. Under his leadership Alelo has developed into a major producer of AI-driven learning products focusing on communication skills. Alelo has developed courses for use in 100 countries around the world, all using the Virtual Role-Play method. Dr. Johnson holds a B.A. in linguistics from Princeton University and a Ph.D. in computer science from Yale University. He is an internationally recognized leader in innovation for education and training. In 2013 he was keynote speaker at the IASTED Technology Enhanced Learning Conference and was co-chair of the Industry and Innovation Track of the AIED 2013 conference. In 2014 he was keynote speaker at the International Conference on Intelligent Tutoring Systems, and was Distinguished Lecturer at the National Science Foundation. In 2015 he was keynote speaker at the ACT Insight Analytics and Emerging Technologies Symposium. In 2017 he was co-winner of the Autonomous Agents Influential Paper Award, and in 2018 he was Distinguished Speaker at IBM Research. When not engaged in developing innovative learning products, Lewis and his wife Kim produce Kona coffee in Hawaii.

Dr. Andrew Naber Dr. Andrew Naber's expertise and research interests focus on individual and team training and performance, leadership, culture, and workforce development policy more broadly. From this research, Andrew and his colleagues developed individual and team training systems, job and decision aids, and measures of performance that support missions, leaders, and teams. Andrew's work has been published in *Human Factors, Journal of Applied Social Psychology, Human Performance*, and the *Encyclopaedia of Industrial Organizational Psychology*. Andrew has worked in both internal and external consulting capacities, and across local government, non-profit, military, and

educational settings. Prior to joining Aptima in November 2016, Andrew was a behavioural scientist at the RAND Corporation. Andrew holds a Ph.D. in Industrial-Organizational Psychology from Texas A&M University.

Dr. Alan Carlin Dr. Carlin is a Principal Research Engineer with Aptima. His interests focus on problems of multi-agent planning and artificial intelligence. These include problems of decision-making under uncertainty, communication between members of a team, meta-reasoning among team members, and multi-agent anytime algorithms. His publications include works on Decentralized Partially Observable Markov Decision Processes (Dec-POMDP); knowledge representation in classical plans; and communication under uncertainty. He has also designed, built, written, and tested hardware and software systems for infrared (IR) and radio frequency (RF) sensors, for use in flight tests. Dr. Carlin received a Ph.D. in Computer Science from the University of Massachusetts, an M.S. in Computer Science from Tufts University, and a dual B.A. in Computer Science and Psychology from Cornell University. As part of his M.S., he also completed the MIT Lincoln Scholars Program, sponsored by the Massachusetts Institute of Technology.

Jared Freeman, Ph.D., is Chief Scientist of Aptima. His research concerns instructional strategies and technologies that accelerate training in complex and ill-defined domains. Dr. Freeman is the author of more than 125 articles in journals, proceedings, and books concerning these and related topics. He holds a Ph.D. in Human Learning and Cognition from Columbia University.

Brenda Lindsay Ms. Lindsay is a Principal Project Manager at Alelo Inc. Ms. Lindsay is project manager of the Alelo Enskill project and has collaborated with DLNSEO on several VCAT language and culture products. She also provided management support on the Tetum project for the Australian Defence Force to provide language instruction for service members deploying to East Timor. Prior to coming to Alelo, Ms. Lindsay worked in the educational software field for nine years. She holds an MBA from Pepperdine University and a Project Management Professional (PMP) certificate.

Subcommittee: Training
Keywords: adaptive training, language, culture, interactive training

"The views expressed do not reflect the official policy or position of the Department of Defense, or the U.S. Government."

Initial Evaluations of Adaptive Training Technology for Language and Culture

W. Lewis Johnson, Brenda Lindsay
Alelo Inc.
Los Angeles, CA
ljohnson@alelo.com
blindsay@alelo.com

Andrew Naber, Alan Carlin, Jared Freeman
Aptima
Woburn, MA
anaber@aptima.com, acarlin@aptima.com
jfreeman@aptima.com

INTRODUCTION

The Department of Defense (DoD) requires large numbers of personnel with sufficient proficiency in foreign languages, regional expertise, and cultural competency to carry out their assigned missions (Work, 2016). Many missions and tasks require at least a limited working proficiency in foreign languages, and some require much higher levels of proficiency. Training personnel to the required level of proficiency, and ensuring that trainees maintain that level of proficiency over time, has been a challenge. Over 70% of the language-coded billets in the DoD are filled by personnel who lack the language skills required for that billet (Freeman & Cohn, 2015).

There are several reasons for this problem. Military language courses are often short in duration, making it difficult for learners to reach the desired level of proficiency in the time available. For example language programs at the US Army Special Warfare Center and School (USAJFKSWCS) are only six months long, and most learners are unable to achieve limited working proficiency (level 2 on the ILR or Interagency Language Roundtable scale) (Johnson et al., 2016). Active duty personnel typically experience extended periods of time in which they have little opportunity to practice their language skills, which leads to language skill decay.

Intelligent tutoring systems and adaptive instructional methods achieve better results overall than conventional instruction (Kulik & Fletcher, 2017) and offer promise as a way to accelerate learning (Sottilare & Goodwin, 2017). In this paper, we investigate how to employ adaptive training technology to achieve better outcomes for language learning. We developed an adaptive training method that offers learners opportunities to practice in realistic situations, automatically identifies gaps in learners' language competencies, and recommends personalized exercises to overcome those gaps. Such an approach offers significant advantages over conventional methods that emphasize passive language skills (reading and listening) and basic knowledge of vocabulary and grammar without realistic practice in using the language. The method and technology is not limited to initial language learning, and could also be applied to competency-based training in other domains.

We have implemented this approach in ALLEARN (Adaptive Language LEARNing), an open architecture for adaptive learning and retention of language skills. ALLEARN provides a simulation-based training environment for practicing language use and assessing language competencies, and recommends personalized learning trajectories that focus on overcoming each learner's individual competency gaps. It is designed to be used in a blended learning curriculum, to help students learn more efficiently outside of class so that classroom time also becomes more efficient and productive.

This article describes the training methodology underlying the ALLEARN environment, and compares it with other methodologies. It then describes the ALLEARN software architecture, system, and technical approach to adaptive training. Finally, it presents results from a formative evaluation of ALLEARN with students and instructors at USAJFKSWCS. The study compared ALLEARN against a non-adaptive version of ALLEARN in which learners followed a fixed curriculum sequence. The study investigated a variety of outcomes between these conditions, such as improved performance on oral and written examinations, time-on-task, and training usage behaviors in the context of blended learning. The findings offer implications for best practices of competency-base, adaptive training, and language learning technology.

ALLEARN TRAINING METHODOLOGY

In an earlier paper, we presented research that identified critical needs for military language learners, both in military and civilian language classes (Johnson et al., 2016). Although these learners are motivated to succeed at learning to advance their careers, they have little free time to study and so must make the best use of available

study time. They also have few opportunities to practice outside of class, so technology that supports language practice and provides feedback would be highly beneficial. Students get practice opportunities and feedback in class, but only if the classes are small. Instructors must spend time helping students who are encountering difficulties, which can slow progress for the entire class.

Students need technology to practice foundational language skills such as grammar and vocabulary, but they also need practice using language in realistic contexts—reading, writing, listening, and speaking. Spoken proficiency was the most important objective for these students, especially the students at USAJFKSWCS who were training to use spoken language in overseas deployments.

The challenges that these language learners face are similar to the challenges that many learners face in competency-based training programs, where the end goal is not academic knowledge *per se* but usable skills that can be applied in real life. For example it is not enough for learners to recognize verb conjugations or demonstrate their knowledge in grammar drills; they must be able to produce grammatically correct utterances in real-time during spoken conversations. Spoken proficiency is particularly challenging because it requires *cognitive fluency* (Segalowitz, 2010)—the ability to comprehend and produce language in real time with a minimum of effort. In this respect the challenges are language training are similar to those in other training domains where skills must be practiced to the point where they are rapid and effortless.

With these considerations in mind, the ALLEARN training methodology follows an adaptive, blended-learning approach. Training modules are designed as interactive supplements that trainees can complete in their own time, on a mobile device. Each targets a competency involving language use in context, for example, buying a plane ticket or getting to know colleagues at work. That way when learners master these modules it should be easy for them to transfer their learning to real-world situations. Because learners can achieve mastery outside of class they can come to class better prepared, making classroom instruction more efficient and productive.

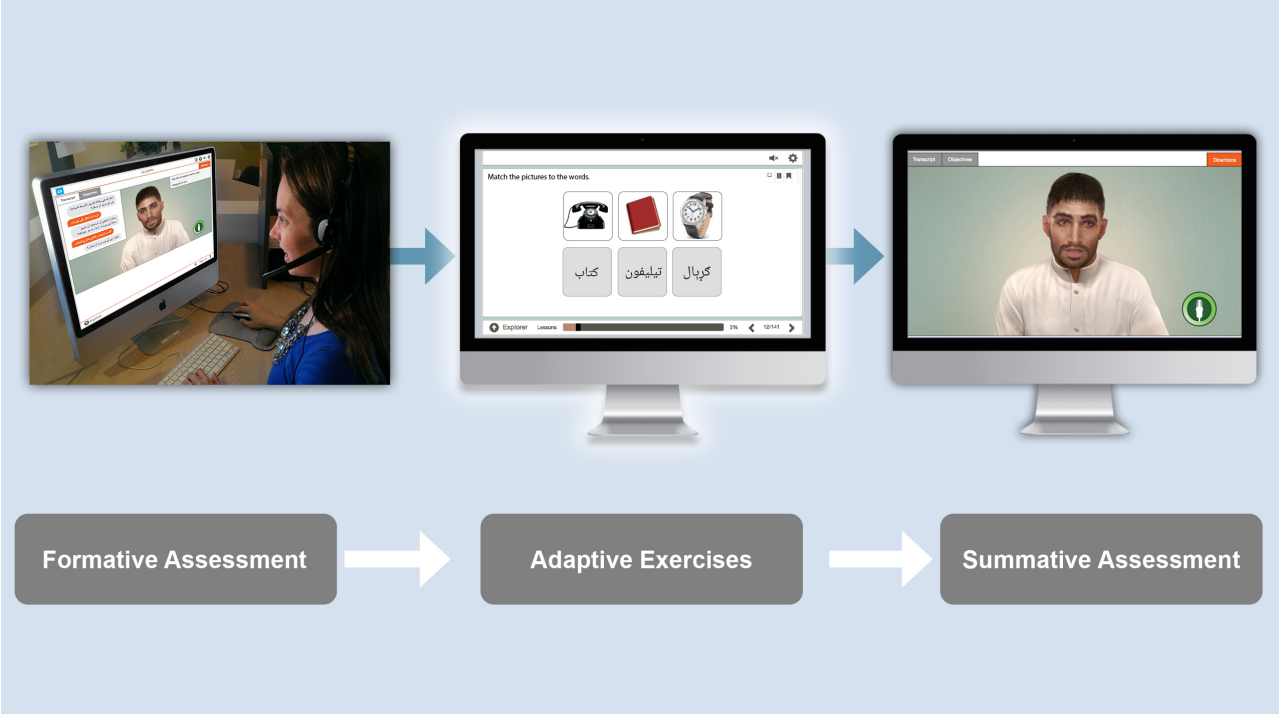

Figure 1. ALLEARN Instructional Process Flow

Figure 1 shows the overall process flow within an ALLEARN training session. Learners start with an interactive simulation in which they converse in the target language with an on-screen character. Speech recognition and natural language processing technology is employed to make the conversation *unscripted*—learners must decide for themselves what is appropriate to say in the situation, rather than follow a fixed dialogue script. If learners encounter difficulties then hints and other scaffolding is available. This serves as a realistic formative assessment to determine whether the learners have mastered the necessary language competencies to complete

the conversation, and if not which language competencies the learner still needs to master.

Once ALLEARN has identified language competencies that the learner needs to develop it automatically selects Learning Units (LUs) consisting of interactive exercises focused on those language competencies. Some of these LUs focus on grammatical competencies (e.g., past tense verbs, nouns and adverbs, prepositions) and some on discourse topics (e.g., travel, education, time residence) Each category includes exercises at one of four levels of difficulty—Easy, Medium, Hard, and Mastery. Easy exercises involve basic recognition and recall of vocabulary and grammatical forms. Medium exercises apply vocabulary and forms in structured exercises. Hard exercises require learners to understand language in context, and Mastery exercises require learners to both understand and produce language in context. Thus, as learners progress from Easy to Mastery exercises they develop mastery of the corresponding language forms. Finally, when learners are ready they can try another conversational simulation, which serves a summative assessment to determine whether the learner has mastered the necessary skills. If not, ALLEARN identifies which competencies need further practice and the cycle repeats.

A key aspect of the ALLEARN approach is the automated selection of exercises at the appropriate level of difficulty to help learners progress toward mastery. We use an adaptive training algorithm that selects exercises to present to learners based on an estimate of the learner's level of mastery of each target skill, referred to in ALLEARN as the *student state*. The algorithm is intended to get learners to a mastery level as rapidly and efficiently as possible. Each learner receives just the training they need, without wasting time practicing skills that they have already mastered.

Note that this methodology is appropriate for refresher training as well as initial instruction. Learners can go back and practice the simulation exercises at any time. If their skills remain sharp, they will complete the simulation easily in just a few minutes. If they encounter difficulties, ALLEARN will select exercises that focus just on those competencies that require reinforcement.

COMPARISON WITH OTHER TECHNICAL APPROACHES

Solutions	Vocabulary learning	Reading & listening	Automated feedback	Conversational practice
Quizlet	✓	x	x	x
GLOSS	x	✓	x	x
Rosetta Stone	✓	✓	✓	x
ALLEARN	✓	✓	✓	✓

Table 1. Comparison of Technical Approaches

ALLEARN differs from other language learning technologies in that it challenges learners to engage in realistic unscripted conversations, uses these conversations as a context for automatically assessing learner competences, and provides adaptive training to help learners master those competencies as efficiently as possible. This combination is unique and distinguishes ALLEARN from other products and methods.

The most common instructional method used in language learning software is flashcard memorization, exemplified by products such as Quizlet (https://quizlet.com) and Memrise (https://www.memrize.com). These are designed to promote recognition and recall of vocabulary items, often out of context. Such memorization exercises could be incorporated into the ALLEARN framework as exercises at the Easy level. These products do not provide training in actual use of a language, i.e., understanding and producing language in context. They can play only a limited role in helping learners achieve cognitive fluency and proficiency.

Products such as GLOSS (https://gloss.dliflc.edu/) include exercises that present realistic examples of written and spoken language, and are designed to promote reading and listening proficiency. Learners typically listen to or read a passage and select from a set of multiple choice responses. Such exercises could be incorporated into the ALLEARN framework as exercises at the Medium and Hard level, if they are aligned with specific language competencies. Some could be used as summative assessments in the ALLEARN framework to confirm that learners have developed speaking and listening skills. But since they do not give learners opportunities to

practice producing language they can at best play a limited role in promoting spoken proficiency.
To help learners promote spoken proficiency spoken language technology is needed to analyze and assess the spoken language that the learners produce. In products such as Transparent Language's CL-150 (https://usg.transparent.com) or McGraw Hill Education's Conéctate the burden of assessing learner speech falls on the instructor or on the learners themselves. For example learners record their own speech and then compare it to a recording of a native speaker. Such approaches are burdensome for instructors and language learners have limited ability to accurately assess their own language performance. In order for an adaptive training approach to apply to spoken proficiency more sophisticated spoken language technology is required.

Our implementation of the ALLEARN architecture utilizes Alelo's Enskill® spoken dialogue platform (Johnson, in press), which lets learners practice unscripted dialogue and automatically assesses language competencies. Enskill builds on previous work at Alelo on role-play simulations, in Tactical Language (Johnson, 2010), VRP MIL (Johnson, 2015), and other related projects (Johnson, 2012). It meets the requirements of ALLEARN for automated formative assessment and summative assessment of spoken language skills and adaptive exercises that involve spoken language production. It is currently being used widely in over twenty countries to practice spoken English (Johnson & Koffler, 2018). Just one organization, Laureate International Universities, has a population of approximately 1,000,000 learners and is providing Enskill to all of their students studying English. Other products that employ speech technology fall short of the requirements of ALLEARN for the following reasons.

- Products such as Carnegie Speech (Eskanazi et al., 2007) have learners read a phrase that is printed on the screen and get feedback on the quality of their pronunciation. Such exercises do not give learners an opportunity to produce language on their own, which is necessary to achieve spoken proficiency. They put too much emphasis on the quality of the learner's pronunciation and not enough on the learner's ability to communicate, express ideas, and convey meaning.
- Rosetta Stone (Pellom, 2012) and DuoLingo (Johnson, 2013) let learners practice scripted conversations, where learners are presented pre-authored choices which they either select or read off the screen. This can be helpful at early stages of language learning when learners rely heavily on memorized phrases, but is less useful for developing productive language skills and achieving higher levels of language proficiency.

ADAPTIVE TRAINING METHOD

The approach to adaptive training used in ALLEARN builds upon Bayesian Knowledge Tracing approaches (BKT; Anderson 1995) and Item Response Theory approaches (IRT; Lord 1980), and POMDP (Partially Observable Markov Decision Process) approaches (Smallwood & Sondik, 1973). BKT uses a Hidden Markov Model (HMM). In the BKT formulation individual components of knowledge are tracked as either learned or unlearned, and a variable called state, which we designate S_n, where n identifies the skill being learned. BKT defines transition probabilities (probability that a student transitions from unlearned to learned), guess probabilities (probability that a student guesses the correct answer to a question), and slip probabilities (probability that a student will make a mistake, even though the student has learned the skill). IRT does not typically include transition probabilities (it does not address the problem of how skills change over time), but it embellishes on the slip and guess probabilities by modeling them with an equation. For example, the simplest 1-parameter form of IRT, referred to as the Rasch model is:

$$p(correct) = \frac{1}{1 + e^{d-\theta}}$$

Where d is the item difficulty and θ is the skill level of the student. POMDP approaches for training were introduced by Levchuk et al.; in this work the authors introduced multiple training actions that were each able to train team skill (Levchuk, 2007, 2012). Each training action is associated with a different transition probability. The team was modeled as being in a high-skill, medium-skill, or low-skill state. The high-skill state was associated with a reward. Unlike the BKT approach above, in this work the model parameters were supplied by Subject Matter Experts. POMDPs address the problem of action selection, to select the action that maximizes reward over time. Formally, the components of a POMDP model are S (a finite set of states), A (a finite set of control actions), Z (a finite set of observations), $\tau(S \times A \times S)$ (the state transition function), $O(Z \times S \times A)$ (the observation function for each action), $R(S \times A)$ (a reward function for each state and action), γ (a discount factor over future time steps), and $b_0(S)$ (an initial distribution that assigns a probability to each state, referred to as a belief state, at time zero).

As in an HMM, a POMDP updates its assessment of student belief state after each action.

$$\Pr(s_1 | s_0, a_0) \sim \Pr(s_0)\, \tau(s_0, a_0, s_1)$$

The approach was matured in subsequent works (Carlin et al. 2013, 2018), to define parameters so that states represent skill attainment over multiple skills. For example, a state of $< A = 1, B = 3, C = 2 >$ would represent that the student has attained levels 1, 3, and 2 on skills A, B, and C respectively). Furthermore, actions represent lesson selection, observations represent an item response function, and reward represents the state multiplied by a weight vector to prioritize selected skills over other skills.

ALLEARN SYSTEM ARCHITECTURE

We have developed and tested a prototype implementation of the ALLEARN system using the approach described above. The implementation has been tested with sample learning content for Americans learning Modern Standard Arabic (MSA). The framework is language-independent and components have also been tested extensively with English content developed for native speakers of Spanish, Portuguese, and other languages.

Figure 2 shows the system architecture for ALLEARN. At the front end, a learner or instructor interacts with the Enskill player, an HTML5-compliant cross-platform Web interface that incorporates speech recognition technology and supports multiple learning activities. These activities enable learners to practice the spoken language skills and assess their performance. ALLEARN provides an application program interface (API) to input activity recommendations and output learner performance analytics. This makes it possible to track learner performance and adapt the sequence of activities to optimize learning. Both the Enskill player and the analytics API support the LTI (Learning Tools Interoperability) v2.0 standard (IMS, 2016). This makes it possible to integrate standards-compliant learning content from a variety of sources.

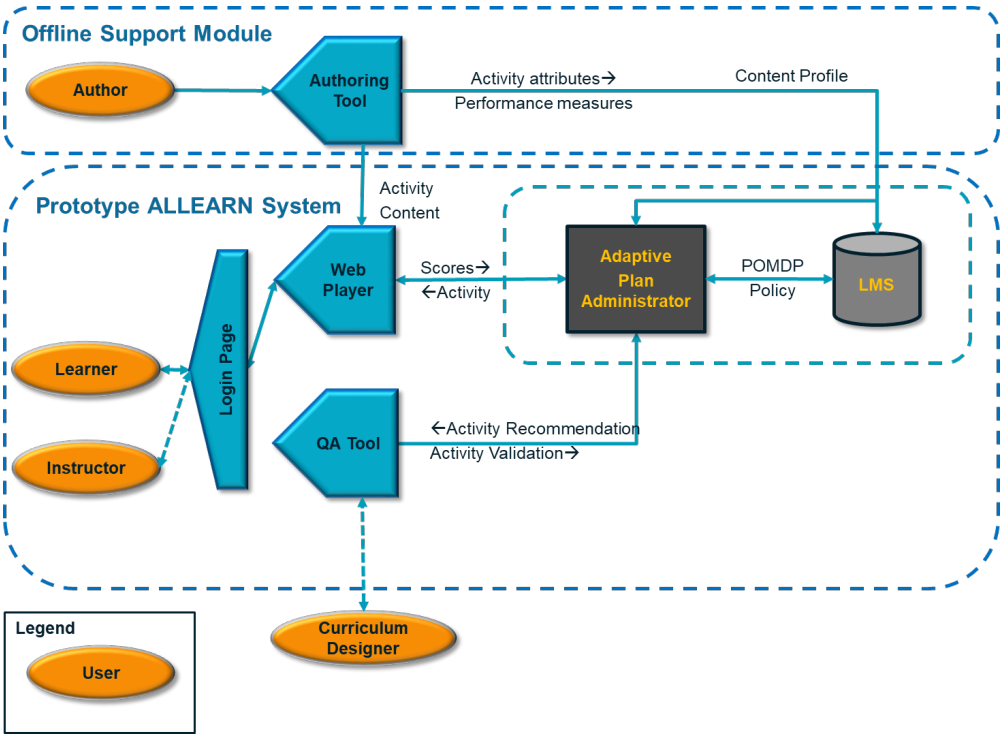

Figure 2. ALLEARN System Architecture

Authors use the Xonnet authoring tool to specify content in a device-independent way, for delivery using the Enskill player. Content authored using other tools is run using the LTI launch mechanism.

On the back end of the system is the Adaptive Plan Administrator (APA). The APA performs two functions. First, it tracks learner state, a report of student skill level over time. After every LU, the results of the exercises (i.e., performance measures) are sent from Enskill to the APA, and the APA updates its assessment of student state. Second, the APA recommends LUs to the student, these recommendations are relayed to the Enskill web

player. The APA saves its results to a learning management system (LMS), which allows for storage of student results, and tracks the state of each student. Thus, multiple students can use ALLEARN concurrently.

In the process flow shown in Figure 1, the ALLEARN system first performs a formative assessment using the Enskill front end. Results of the formative assessment are relayed to the APA back end. The APA assesses student state and recommends an LU. Enskill then delivers this LU to the student (a Learning Unit is composed of several Exercises; each Exercise contains one or more correct/incorrect measures), and when the LU is finished Enskill relays the exercise scores to the APA. The APA updates its assessment of student state and issues a new LU recommendation. The process repeats as the next LU is delivered.

The final element of the system is a QA tool for the back end. The QA tool supports validation and improvement of the adaptive trainer by a human. Validating training recommendations presents two challenges. First, unlike typical software, where a "bug" is obvious and rules based, and can be described as a behavior, a "bug" in an adaptive training recommendation is concept-based, not as easily specified, and in most cases can only be found by a human. For example, a bug in an adaptive trainer can be "the student was recommended to do a Future Tense Exercise at High difficulty, despite performing only so-so on a recent Nouns and Verbs question at Low Difficulty which used the future tense". Second, it is time-consuming and requires knowledge of the target language, e.g., Arabic, to QA the tool by running through the actual exercises. The ALLEARN QA tool allows a user to interface to the back-end directly, without necessarily taking the lessons. The QA tool shows student state (assessment with respect to the 19 Arabic language competencies included in the prototype). The user then uses the QA tool to simulate that student taking one of the LU's. The user selects the LU to be taken (usually this will correspond to the training recommendation provide by the APA), then inputs the simulated scores on all of the exercises in the LU, and presses a "Submit" button to send the simulated result to the APA. The APA then updates its assessment of student state, displayed on a UI, and also issues a training recommendation.

ADAPTIVE LEARNING STRATEGY CONSTRUCTION

As mentioned previously, the adaptive strategy performs two functions. First, after LU results are received the adaptive learning strategy updates its assessment of student state. Second, it maps student state to a recommendation for the next LU. Chaining these two functions together, this means that after learning results are received, the strategy converts LU results to LU recommendations. Because of this, the strategy is adaptive, every unique path of results leads to a unique path of recommendations. In the rest of this section, we outline the steps that were taken to construct the adaptive learning strategy. Each step below is given a dual label, consisting of what the SME does, along with what component of the POMDP model it was used to build.

Identify Domain Skills/Define Student State. The APA tool has been used in other domains; in the language domain the first step was to identify the training objectives and the set of skills that were part of this domain. Practitioners in different domains speak about these terms in different ways, and the lexicon differs between disciplines. The ALLEARN adaptive training software requires only a simple list as inputs; we decided to describe the items in that list as Language Competencies (LCs). The LCs used in ALLEARN prototype were: Past Tense Verbs, Present Tense Verbs, Future Tense Verbs, Prepositions and Adverbs, Questions, Travel, Education, Measurement, Negation, Numbers, Advanced Numbers, Time, Residence, Nouns and Adverbs, Family, Word Order, Age, Possessives, and Calendar. We also decided to track student state on a 0-4 scale. These decisions together were sufficient to define POMDP state. A student's state consists of the student level for each of the 19 LCs above. That is, an example of a state the student could be in was, "Past Tense Verbs=0, Present Tense Verbs=1," etc.

Label Content/Construct Transition Probabilities. The Enskill player presents LUs to the student, which comprises several exercises. For instructional reasons, authors recommended that exercises not be recommended separately, only as part of the LU. However, different exercises could cover slightly different LCs. For example, an LU on "Time" consists of multiple exercises involving time of day, but one exercise may additionally generate a measure of Present Tense Verb skill (ask the learner to label the time of day) while a different exercise may generate a measure of Future Tense Verb skill (ask the learner to make an appointment).

Thus, as a novelty from existing literature, the adaptive learning strategy needed to consider two different granularities. Furthermore, as mentioned above, the adaptive learning strategy needs a probability of transition between each pair of states, given an LU (e.g., probability of transitioning from Level 1 to Level 2 in the Past Tense Verbs skill, for each LU). These types of probabilities are daunting for Subject Matter Experts to specify by hand.

To address these two challenges, we asked subject matter experts to specify for each Exercise two pieces of

information: (1) The LCs relevant to the exercise, as well as how relevant those LCs were to answering the exercise (very relevant or somewhat relevant), and (2) the difficulty of the exercise. The information from the exercises was aggregated into an overall Applicability and Difficulty score for each LU. Then POMDP transition probabilities were built using the Applicabilities and Difficulties, using the following rubric:

- The higher the applicability, the more likely the LU moves the student to a higher level in that LC.
- The closer the match of difficulty to student state, the more likely the LU moves the student to a higher level in that LC.

The equation used was:

$$\boldsymbol{\tau(s, < a, d >, s') \sim \frac{(3|d - s| + 1)(s' - s)}{a(L_{\{max\}} - 1) + .01}}$$

Where $\boldsymbol{d}$ and $\boldsymbol{a}$ represent the difficulty and applicability of the LU, $\boldsymbol{s}$ and $\boldsymbol{s'}$ are the old and new state levels of the learner on that LC, and $\boldsymbol{L_{\{max\}}}$ is the maximum possible level (4). The sum of transition probabilities is normalized.

Generate Observation Probabilities. Observation probabilities were also specified for each learning unit. Each LU was decomposed into the exercises within. For example, a student who answered 6 exercises correctly and 2 exercises incorrectly was interpreted as generating 8 observations, with the equations governing each observation governed by the difficulty and applicability of that exercise. The observation function used was from Item Response Theory in the previous section. In that equation, the probability of correctness was known (1 if correct, 0 if incorrect), as was the exercise difficulty, leaving the only unknown as $\boldsymbol{\theta}$. After each LU, for each exercise within, the likelihood of each possible integer value of $\boldsymbol{\theta}$ was computed, and then the overall distribution was normalized to sum to one.

Generate Reward Function. A reward function was generated for each state. In general, a reward was accumulated for each LC as the student level for that LC. (e.g., a student who is a "3" at Past Tense" is assigned a reward of 3 for that state in the POMDP model). However, the ALLEARN curriculum was divided into two lessons, each lesson had its own set of LCs as objectives. LCs that were lesson objectives were assigned reward according to the above rubric, whereas LCs that were not part of a lesson were assigned a Reward of zero for all levels within.

Implement Learning Strategy. With the above information defined, after each lesson, the result of all the exercises was sent from Enskill to the adaptive strategy. The strategy would first use the transition probability and observation probabilities to update its assessment of student state on each LC. The new state is referred to as a "belief state", because the new state is a probability distribution over states, not a single state. However, in practice the user interface we used for QA simplified this information by presenting an expected value for each LC (e.g., a student who is 20% likely to be at level 1 for an LC and 80% likely to be at level 2 is displayed as being at level 1.2 on the UI).

After updating the student state, the learning strategy would consider each possible LU. It would invoke that lesson's transition probability, and consider the expected value of the new state after that LU. The LU with the highest predicted expected value would be selected.

To avoid repeating LU, once a lesson was taken, its transition probability for all LCs addressed by the lesson was reduced by 75% if the student passed the LU, and by 25% if the student failed, for each level that was higher than the current level. The transition probability of staying at the same level was increased by the corresponding amount. This modeled a lesson losing its effectiveness if taken more than once. Thus, while it was still possible to repeat lessons when particularly suited to a weak LC, the bar to select repetition became higher and higher with each repetition.

INITIAL RESULTS OF PILOT EVALUATION AND FOCUS GROUPS

During the spring of 2018, a pilot evaluation and focus groups were conducted at USAJFKSWCS with MSA learners to obtain their feedback regarding content design and user experience, and offer insights for future adaptive language platforms. The pilot evaluation employed the following sample and method.

Sample

The pilot evaluation included two cohorts of MSA students, separated into two classes. The pilot evaluation obtained data from five students total: three using a non-adaptive version, and two using an adaptive version. In the non-adaptive version adaptive training was disabled and students followed a pre-specified sequence of exercises. The pilot learning intervention covered the first two lessons of Module 2 of the USAJFKSWCS

curriculum. The USAJFKSWCS curriculum comprises six modules with 4-6 lessons per module. MSA students were provided with login information to access exercises via a website. Because students may not have been able to access exercises with a personal device, they were also provided with a tablet.

Method

The adaptive cohort received exercise recommendations as described previously. The non-adaptive training policy was developed through the following method. A mock student completed all LUs *incorrectly* on the initial formative assessment. Next, the mock student completed all LU's *correctly*. The suggested sequence following this path formed the non-adaptive sequence of learning exercises.

Instructors were encouraged to recommend that students include ALLEARN in their self-study at their discretion, and the researchers recommended ALLEARN usage of about 30 minutes a day. Participants were free to decide for themselves how to divide their time between ALLEARN and other self-study activities.

During the course of the lesson plan students were tested before and at the end of the module. However, given the small sample size and restriction of range for these scores, these were not particularly informative. Instead, the research team focused on (1) identifying whether the adaptive version of ALLEARN offered unique paths to individual students, and (2) students' feedback regarding adaptive language training.

Adaptive Metrics

Students in the non-adaptive version completed 24 unique LUs, whereas those in the adaptive version completed 45 unique LUs. In terms of the proportion of LU's covered, students in the non-adaptive version completed 30% of the available curricula, whereas students in the adaptive version completed 56%. Finally, adaptive students reached the highest difficulty levels (4), in contrast to those in the non-adaptive, that did not surpass the third difficulty level.

In terms of the degree of adaptation, students in the non-adaptive version naturally did not deviate at all from the base sequence; whereas, those in the adaptive version received 22 LU's presented at a different time point in the curriculum from base sequence (46%). Furthermore, those in the adaptive version received eight LUs that were not received by students in the base sequence at all (10% of content).

Focus Group Reactions - Adaptation

In terms of adaptive language training policy and pathing, adaptive lesson sequences violated student expectations of repetitive drill and practice. That is, students reported that they were unsure why exercises' content adapted. Students recommended that additional feedback be provided related to their performance—particularly, in regard to the performance dimensions related to adaptive training recommendations.

MSA students described desirably blended learning tools as focusing on "reps with variation" and vocabulary building. They offered the suggestion that exercises could be a single grammar construction that is repeated multiple times, but with unique vocabulary words substituted in and out to provide variation. Students reported that this would be particularly valuable as MSA contains word families (and gender/plurals) that may be easily grouped into exercises and are particular to MSA.

Focus Group Reactions – Matching Spoken Exercises to Technology

Finally, students found the automated speech recognition technology in conjunction with virtual agent simulations to be unreliable and potentially demotivating due to trust in the technology. Specifically, students reported that they were unsure how to correct their performance during ASR simulations, because they did not receive specific feedback diagnosing their problem. Additionally, not all students realized there was a transcript option available that tracked their MSA speech-to-text, which may have alleviated some issues.

In contrast, students found the recording/playback feature of their own Arabic spoken responses to be very valuable. Accordingly, they suggested retaining features allowing recording/playback, but not formal scoring

using automated speech recognition that may prevent progressing through future lessons. Instead, they reported an override/honor-system instead of an ASR feature, similar to CL-150, would be valuable.

Focus Group Reactions - Blended Learning and Desired Exercises

Complementing adaptive training, students reported that they appreciated access to the catalog of exercises to self-select content areas as well. The adaptation mechanism developed auto-adjusts the learning path, even when students self-select exercises.

In regards to alternate spoken and listening exercises, students reported that they would appreciate listening exercises that include dialogue between *two* native MSA speakers, as they reported that there were few opportunities to observe different MSA speakers' voices in learning materials. In contrast, students reported some frustration with spoken exercises that require students to portray a character, as they struggled to remember details about that character. Second language acquisition and performance already taxes working memory, and it would appear that roleplaying likely increases that cognitive load further (Linck, Osthus, Koeth, & Bunting, 2014). Nevertheless, role-playing activities were frequently used in SWCS classes.

DISCUSSION

This study investigated learner reactions to an adaptive MSA tool designed to be used within a blended learning environment. The study investigated whether the adaptive version actually suggested different paths in learning from the non-adaptive users. Generally, learners reported positive reactions to adaptation but were interested in more feedback about their performance in relation to how adaptive recommendations were made.

Learners reported some challenges with the speech recognition as it related to spoken dialogue with a virtual agent, but reported significant value in opportunities to practice pronunciation and speech. This contrasts with the experience of users of Alelo's Enskill English product, which uses similar speech technology and is evaluated highly by learners (Johnson, in press). ALLEARN leverages Microsoft's Cognitive Speech Services to process and evaluate the learner's speech inputs. An analysis of the speech input results determined that many of the recognition failure were due to learner accents. A review of the audio files indicates that learners were producing the proper responses, but they were being interpreted by the system as incorrect. Enskill English overcomes these problems through the use of a database of common pronunciation errors, extracted from learner speech recordings, and a statistical natural language processing algorithm that is able to recognize the intended meaning of an utterance even when the learner makes mistakes in pronunciation, grammar, or usage. We recommend incorporating similar methods into ALLEARN.

Based on the feedback from the participants we are planning improvements to the user interface to make the simulated conversations easier to use. This includes making the dialogue objectives and dialogue transcript more easily accessible, and providing feedback on learner errors. We are also analyzing the dialogue transcripts and speech recordings to identify further areas for technical improvement of the spoken dialogue system.

Future trials and releases of ALLEARN should include detailed orientation materials for instructors, such as webinars, tutorials, and sample lesson plans. This will make it easier for instructors to blend ALLEARN into their courses. Instructors have an important role to play in helping students get started with the tool and getting as much learning benefit from it as possible.

ALLEARN's adaptive recommendation engine provided the curriculum sequencing of LUs; however, learners were also able to self-select LU's. Generally, students prefer some freedom to decide what LUs to work through, and indeed, this sample reported that they sometimes self-selected exercises. Regardless of whether the recommendation engine or the learner selected an LU, performance information was recorded and used to inform subsequent adaptive recommendations. Learners valued feedback not only on their performance but sought the rationales for the adaptive educational paths. Providing feedback is a straightforward enough instructional approach; however, in the domain of adaptive learning recommendations this issue is more complex and a subject for future research.

Limitations

Some of the conclusions of this pilot evaluation should be tempered by the unique nature of the sample and curriculum. First, MSA morphology is quite complex compared to English, which is the reason that some students gave for why exercises that repeat the same grammar but build vocabulary would be valuable. Second,

the SWCS course is unusual in that instructors are primarily native speakers with limited formal knowledge of MSA grammar. SWCS students reported that they took a lot of time to seek out formal grammar lessons to complement and explain the naturalistic style of instruction used in class. SWCS is an intensive residency program focusing on speaking—an opportunity many students do not have in other language learning settings. Finally, the small sample size limited comparison between the adaptive and non-adaptive curriculum sequences.

Future Research

Adaptive training, particularly in language acquisition, offers a fertile domain for future research. These results offer two viable research streams: (1) improving sequencing and recommendations of adaptive training content and (2) providing more informative feedback and explanation of adaptive recommendations. These two issues are clearly interrelated. Future research could inform development efforts on the effects of an adaptation policy of *massed* practice of similar content (which aligns with student expectations, but may reduce retention) in place of the current policy of *distributed* practice whereby language competencies may seem to switch abruptly (which violates expectations but maximizes retention). Accordingly, students suggested more feedback about both their language performance and *how* it led to adaptive transitions. Additionally, it may be helpful to familiarize students with distributed practice (i.e., frequent "hops" between training topics) and explain why it is advantageous (Arthur, Bennett Stanush, & McNelly, 1998). Optimizing the frequency and variability of hops may not only improve learning outcomes, but also learners' acceptance of adaptive recommendations and feedback. Although ALLEARN could take advantage of the learner's choices to improve its own recommendations, it is not yet clear how these may be more or less informative to adaptive recommendations. That is, if a learner decides that she needs to work on a particular language skill, it may align, operate in parallel, or even be at odds with adaptive recommendations. For example, a learner may be progressing through a particular skill and self-select the next recommended exercises (alignment), another exercise of a different skill (parallel), or a lower level of exercise of the same skill (at odds), respectively. This learner may or may not be appropriately self-monitoring, and adaptive recommendations might revise its own estimate of the learner's mastery based on not only performance information, but also self-selection information.

CONCLUSIONS

This paper presents the design and preliminary evaluation of adaptive training technology for language and culture. Initial results show promise and are informing further iterative development of technology and implementation materials for instructors. Once the digital activities and learning model are fully developed, the research team intends to adapt the Xonnet authoring tool used internally to support authoring by instructors. Ongoing research is validating the ALLEARN approach, to demonstrate that it enables greater numbers of learners to acquire higher levels of spoken proficiency compared to traditional methods.

Our intention is to develop ALLEARN into a model for adaptive language learning, both for civilian education and military training. As major educational publishers transition their language-learning offerings into digital products, there will be increasing demand for language-learning content that adapts to the learner. On the military side, the ALLEARN approach can support learners in the schoolhouse and continue to support them through their military career. It can help military members sustain their language skills during the extended periods of time when they have limited access to language instruction. The adaptive training method has applicability beyond language learning to other domains whether the goal of training is to master key competencies and apply them proficiently in realistic settings.

ACKNOWLEDGEMENTS

ALLEARN is sponsored by the Office of Secretary of Defense (Personnel and Readiness) and the Defense Language and National Security Education Office (DLNSEO), and in partnership with the Office of Naval Research (ONR) and Special Operations Forces Language Office (SOFLO). The authors wish to thank everyone on the ALLEARN team, especially Kent Halverson, Diane Kramer, Curtis Wingert, and Autumn Furnish. We wish to thank our Government sponsors and collaborators, including Peter Squire, Gary Bauleke, Jack Donnelly, Mike Judge, as well as their teams. We also wish to thank our I/ITSEC birddog, John Schlott, for his helpful reviews and comments. This work is funded under contract N00014-16-C-1041. The views, opinions, and/or findings contained in this article are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied, of US Government.

REFERENCES

Arthur Jr, W., Bennett Jr, W., Stanush, P. L., & McNelly, T. L. (1998). Factors that influence skill decay and retention: A quantitative review and analysis. *Human performance*, *11*(1), 57-101.

Carlin, A., Dumond, D., Freeman, J., & Dean, C. (2013). Higher Automated Learning through Principal Component Analysis and Markov Models. In *Artificial Intelligence in Education* (pp. 661-665). Springer Berlin Heidelberg.

Carlin, A., Ward, D., & Freeman, J. (2016). Representation, Selection, and Scheduling of Training in a Lifelong Learning Context. ModSim 2016.

Carlin, A., Oster, E., Nucci, C., Oster, E., Kramer, D., & Brawner, K (2018). Data Mining for Adaptive Instruction. Florida AI Research Society Conference (FLAIRS-31).

Eskenazi, M., Kennedy, A., Ketchum, C., Olszewski, R., & Pelton, G. (2007). The NativeAccent™ Pronunciation Tutor: Measuring success in the real world. In M. Eskenazi (Ed.) *Proceedings of SlaTE Workshop on Speech and Language Technology for Education*. ISCA Tutorial and Research Workshop, Farmington, PA USA.

Freeman, H. & Cohn, J. (2015). S&T for Blended Adaptive Language & Culture Training Symposium. *HPT&B Newsletter* 3, p. 21.

IMS (2016). Learning Tools Interoperability. *IMS Global Learning Consortium*. Retrieved May 27, 2016 from https://www.imsglobal.org/activity/learning-tools-interoperability

Johnson (2013). Review: Duolingo and Babbel. *The Economist*, June 14, 2013. Retrieved May 19, 2016 from http://www.economist.com/blogs/johnson/2013/06/language-learning-software.

Johnson, W.L. (2010). Serious use of a serious game for language learning. *International Journal of Artificial Intelligence in Education*, *20*(2), 175-195.

Johnson, W.L. (2012). Error detection for teaching communicative competence. *Proceedings of the IS ADEPT Symposium*, Stockholm,37-42.

Johnson, W.L. (2015). Constructing Virtual Role-Play Simulations. In R Sottilare, A. Graesser, X. Hu, K. Brawner (Eds,), *Design Recommendations for Adaptive Intelligent Tutoring Systems: Authoring Tools* (Volume 3), 211-226. US Army Research Laboratory, Orlando.

Johnson, W.L. (in press). Data-Driven Development and Evaluation of Enskill English. *International Journal of AI in Education*.

Johnson, W.L., Carlin, A., Freeman, J., & Cohn J.V. (2016). Adaptive Training Technology for Language and Culture. *Proceedings of I/ITSEC 2016*.

Johnson, W.L. & Koffler, R. (2018). *Alelo's Enskill Platform Expands into Sixteen Countries*. Retrieved May 18, 2018 from http://www.prweb.com/releases/2017/12/prweb14976697.htm.

Kulik, J.A. & Fletcher, J.D. (2017). Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. IDA Document NS D-8391. Alexandria, VA: Institute for Defense Analyses.

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review*, *21*(4), 861-883.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Pellom, B. (2012). Rosetta Stone REFLEX: Toward improving English conversational fluency in Asia. *Proceedings of the IS ADEPT Symposium*, Stockholm, 15-20.

Puterman, M. L. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. Hoboken, NJ: Wiley.

Segalowitz, N. (2010). *The cognitive bases of second language fluency*. London: Taylor & Francis.

Sottilare, R. & Goodwin, G.A. (2017). Adaptive Instructional Methods to Accelerate Learning and Enhance Learning Capacity. Barcelona, Spain: International Defense and Homeland Security Simulation Workshop.

Work, R.O. (2016). *Defense Language, Regional Expertise, and Culture (LREC) Program*, DoD Directive 5160.41E, amended Feb. 9, 2016. Retrieved May 17, 2016 from http://www.dtic.mil/whs/directives/corres/pdf/516041p.pdf.