Adaptive Training Technology for Language and Culture

W. Lewis Johnson	Alan Carlin, Jared Freeman	Joseph V. Cohn
Alelo Inc.	Aptima	Defense Health Agency
Los Angeles, CA	Woburn, MA	Falls Church, VA
ljohnson@alelo.com	acarlin@aptima.com, jfreeman@aptima.com	Joseph.v.cohn.mil@mail.mil

ABSTRACT

The Department of Defense has over 40,000 positions requiring foreign language skills. But most of the people in these billets lack the full, required skillset. This has a serious impact on military readiness. There are two main reasons for it. Training is slow, often requiring 6-12 month residency, and most trainees fail to achieve the level of proficiency that military training billets require. After resident training, personnel tend to lose their language skills unless they have opportunities to continually practice them.

This paper reports on progress toward a novel solution called ALLEARN (Adaptive Language LEARNing) to address this challenge. The solution comprises these elements:

- A blended learning approach that combines the expertise of language instructors in guiding and managing learning with adaptive personalized instruction,
- A highly interactive, virtual training environment in which learners apply and assess their language skills, and
- Adaptive training technology and architecture that selects training content in a manner that accelerates learning, manages decay, and accounts for individual differences in learning rates.

Pilot evaluations are being conducted with military personnel seeking to achieve spoken proficiency in Modern Standard Arabic.

To better understand learning needs and current barriers to achieving spoken language proficiency, interviews of a total of 17 instructors and 22 students were conducted at US Army JFK Special Warfare Center and School (USAJFKSWCS) and San Diego State University (SDSU). The interviews highlighted a need for online learning activities that provide realistic opportunities to practice spoken language skills in unscripted settings, immediate meaningful feedback, and automated assessments of communicative competence that reduce the burden on instructors. We have developed a system architecture to address these needs. It is an open architecture for language practice and assessment that collects analytics on learner performance and optimizes learning trajectories

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 innovative learning products focusing on communication skills. Alelo has developed courses for use in a number of 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 2012 he was keynote speaker at the International Symposium on Automated Detection of Errors in Pronunciation Training in Stockholm. 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. When not engaged in developing disruptive learning products, Lewis and his wife Kim produce Kona coffee in Hawaii.

Dr. Alan Carlin, Dr. Carlin is a Senior 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.

Joseph Cohn, PhD., is an Aerospace Experimental Psychologist in the U.S. Navy's Medical Service Corps. His research interests lie at the intersection between Human Systems and Biomedical Research. He has directed Basic, Applied, Advanced Technology Development, Advanced Component Development and Prototype programs developing performance-enhancing biomedical and human systems technologies for the Army, Navy, Air Force and Marine Corps.

Adaptive Training Technology for Language and Culture

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

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

Joseph V. Cohn
Defense Health Agency
Falls Church, VA
Joseph.v.cohn.mil@mail.mil

INTRODUCTION

Foreign language skills, regional expertise, and cultural competency are essential in the missions of the Department of Defense (DoD) (Work, 2016). The deployments to Iraq and Afghanistan heightened awareness of this issue yet the DoD has struggled to develop and maintain language skills within the Total Force. According to a recent report to Congress on foreign language proficiency, the DoD is able to fill less than 30% of its ~41,000 language-coded billets with personnel having the required proficiency in the language specified (Freeman & Cohn, 2015).

The situation in US Special Operations Command (USSOCOM) illustrates this challenge well. Language skills and cultural awareness are foundational for Special Operations Forces (SOF), and essential for building rapport, maintaining situational awareness, and communicating effectively with foreign partners (Donnelly, 2015). A needs assessment of US Army Special Forces language proficiency requirements (SWA Consulting, 2010) determined that of 212 SOF language-related tasks, 167 required an ILR (Interagency Language Roundtable) spoken proficiency level of 2 or above (i.e., an ability to handle limited work-related interactions) to be fully effective. Yet only 30% of US Army Special Forces (USASOC) will be qualified at level 2 if USASOC achieves its current goals. Currently 36% of graduates from the 6-month intensive language program at the US Army JFK Special Warfare Center and School (USAJFKSWCS) score at level 2 or above on an oral proficiency test, and only 25% of the Arabic language graduates do so (A. Nabipour, personal communication). Once students leave the school they have limited time and opportunity to practice their language skills. Most of those who managed to achieve a proficiency level of 2 in school revert to a level of 1 or 1+. Learners who fail to achieve level 2 proficiency are at even greater risk of skill decay. This pattern of decay is consistent with research on language school graduates at the Defense Language Institute Foreign Language Center (Shearer, 2013).

The ALLEARN (Adaptive Language LEARNing) project is conducting research to address the twin challenges of acquisition and retention of language skills by operational forces, under sponsorship of 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 and Special Operations Forces Language Office (SOFLO). The goal of the project is to develop advanced learning technologies and methods that produce higher levels of language proficiency and retention over time. Although the long-term goal is to support rapid learning of all four language modalities–listening, speaking, reading, and writing–the primary focus of this initial phase of work is on spoken proficiency. Spoken proficiency is essential for SOF, and important for all military members deployed overseas.

We hypothesized that the following problems might be preventing learners from developing spoken language skills rapidly and retaining them.

- Students lack access to opportunities for practicing spoken language skills and receiving feedback outside of the classroom. As a result, valuable classroom time is devoted to practice activities for developing basic spoken language skills.
- Students in language classes vary greatly in skill level and learning rate. Relatively large and diverse classes make it difficult for instructors to address the needs of individual needs of learners.

Adaptive, computer-based language instruction could increase the opportunities for practice out of class and in it, focus practice on weak skills, and "buy time" for advanced instruction in class. Consequently we envision that a complete system should encompass the following elements:

- An instructional strategy of blended learning, which combines the expertise of language instructors in guiding and managing learning with adaptive personalized instruction;

- A highly interactive, virtual training environment in which learners apply and assess their language skills; and
- Adaptive training technology and architecture that selects training content in a manner that accelerates learning, manages decay, and accounts for individual differences in learning rates.

This paper summarizes progress of the ALLEARN project to date. It is organized as follows. First, it summarizes the iterative design methodology that we are employing to ensure that the solution addresses the needs of language instructors. It presents findings from an in-depth analysis of the needs of military language learners and their instructors. We then provide a description of the system architecture for ALLEARN, and discuss how the system is intended to address these needs. Finally, the paper summarizes next steps in the project and the research questions that they are intended to answer.

RESEARCH METHODOLOGY

This work applies the approach of design science to discover, develop, and test language training solutions. This approach, defined by Collins (1992) and implemented by Brown (1992), engineers training technologies and tests them in cycles that exploit the rapid development methods of modern software engineering, and the experimental perspective of instructional science. Specifically, second-language-acquisition (SLA) experts, instructors, and learners at DLNSEO-supported institutions participate in the creation and formative evaluation of learning materials through interviews and focus groups. We will augment expert judgment, in 2017, with educational data mining to discover efficiencies that can further accelerate learner movement to higher levels of proficiency. The team then applies Agile development techniques to design, develop, and deliver successive versions of the system. Evaluations of each delivered prototype are conducted in instructional settings. They include standard usability analyses as well as instructional research that tests the efficacy of the approach relative to current classroom instruction. These experiments are designed to identify the features and content of ALLEARN that efficiently augment or replace basic classroom instruction. Discoveries from these experiments guide the next iteration of design, and should eventually produce a system that provides students with focused practice, and instructors with relief from instruction in language basics, so that they can focus on more advanced language skills. The combination of human instruction and instructional technology will constitute a blended learning environment for efficient language learning.

Under the advice of the SOF Language Office, the project is currently focusing on Modern Standard Arabic (MSA). There is an enduring need for special operators with Arabic speaking skills, and students often struggle with Arabic because it is a Category 4 language, the most difficult category for English speakers. Although each Arabic country has its own colloquial dialect, MSA is understood throughout the Arab world and provides a good foundation for military members deploying to Arabic-speaking countries.

NEEDS ANALYSIS

The initial phase of work consisted of a set of focus-group interviews of instructors and students of Arabic, to understand their needs and the challenges they face. Interviews of instructors took place at San Diego State University in April 2016, and interviews of SDSU students took place in July 2016. Interviews of instructors and learners were conducted at USAJFKSWCS in Ft. Bragg, NC in May 2016. Both programs are intensive resident language programs for military personnel. The SDSU program caters to a broad range of military personnel, while the USAJFKSWCS program focuses on US Army personnel seeking to qualify as Special Forces, aka Green Berets. We also observed classes in action, to understand what learning activities instructors employed in the classes and note the language skills and skill deficits of the learners.

A total of 17 instructors were interviewed: 5 at SDSU and 12 at USAJFKSWCS. Instructors varied greatly in terms of teaching experience. Many had been teaching Arabic for years. Some of the instructors at USAJFKSWCS were native speakers of Arabic but had little or no training in language teaching.

A total of 22 students were interviewed: 11 at SDSU and 11 at USAJFKSWCS. Students were young adults, ranging in age from 18 to 25. Classes at USAJFKSWCS included a mix of officers, typically with a college education, and enlisted personnel without a college education. Classes at SDSU consisted of young military personnel who were learning Arabic to meet a professional training requirement and a few civilian students who were interested in Arabic for personal fulfillment.

Main Findings

Our interview data was analyzed based on what was most important to the students and what was most frequently mentioned as part of the student learning experience. The following topics were mentioned most frequently and emphatically during the student interviews.

Motivated, overworked learners. These learners were motivated to succeed in learning the language, mostly to advance their careers and receive bonus pay. But they must cover a large amount of material in a very short amount of time, and some learners find the pressure to be daunting. They have little free time in which to study, and must squeeze in language learning among their many other training and home obligations. They therefore need tools they can use whenever they have a few free moments to practice.

Limited knowledge of grammar and multilinguistic awareness. Many students have a very limited understanding of grammatical concepts, and are unable to identify a noun or a verb in an English sentence. This poses a challenge for learners of Arabic, a language with complex grammatical forms, many of which are very different from English. The instructors believed that students with a poor understanding of grammatical structures were less likely to achieve Arabic proficiency. These observations are consistent with research findings in the literature that low-proficiency L2 students tend to lack metalinguistic awareness (e.g., Tokunaga, 2010).

Focus on the outcome: Spoken proficiency. The primary objective of most of these courses is to help the learners achieve significant gains of spoken proficiency, as measured using an Oral Proficiency Interview (OPI). The one exception was an SDSU class of professional military linguists who were mainly focused on scoring well on the Defense Language Proficiency Test (DLPT), a reading and listening test. The emphasis on spoken proficiency has an impact throughout the curriculum, especially at USAJFKSWCS. Instructors conduct mock OPI assessments with learners over the phone, to help get learners comfortable with oral proficiency testing. However instructor assessments of proficiency are subjective and the reliability of these assessments is unknown; this makes it difficult to track learner progress with confidence.

Importance of reading, listening, and writing skills. Although the primary focus of instruction was on spoken proficiency, the instructors felt that reading, listening, were important enablers for achieving spoken proficiency. Some felt that basic writing skills were essential as well. Students needed these skills in order to take advantage of Arabic language resources available over the Internet.

Class size and diversity of language skills within a class significantly slowed progress. Instructors typically are forced to devote their attention on the learners who are having difficulties, instead of helping the more advanced learners. This is an issue even though the class sizes are relatively small, ten students at SDSU and six at USAJFKSWCS.

Role-play is a major element of classroom activities. Instructors divided learners into small groups and have them role-play various conversational exchanges. This gives the learners more confidence to speak.

Technology is needed for practicing spoken language skills. The instructors wanted simulated role-plays similar to the in-class role-play activities. This would let learners practice speaking skills on their own, in a safe environment, and get immediate feedback. They wanted to see learning activities that incorporated automated speech recognition in Arabic. Some students are shy about speaking in a group setting, or less eager to participate in class discussions, and so computer-based practice would help. It would be most helpful if the responses of the non-player characters in the role-plays were unpredictable, so learners could not rely on memorized scripts to complete the exercise. Novice language learners tend to rely heavily on memorized language, and they must get beyond this if they are going to advance to higher levels of proficiency. Such a capability would also give learners opportunities to practice speaking with various types of people. For example, students at USAJFKSWCS are all male, which means they have few opportunities to practice speaking with females.

Technology is needed for practicing foundational skills. These include Arabic pronunciation and the Arabic alphabet. Difficulties with these skills impede development of spoken proficiency.

Instructors and students want to create content. Both instructors and students were keenly interested in creating their own content. Some of the students at USAJFKSWCS attempted to create their own study materials such as vocabulary lists, but found that the existing tools for creating them were not user-friendly enough to maximize their study and were not portable to mobile devices.

Discussion

These findings support the hypothesis that there is a significant need for tools for practicing and assessing spoken language skills. These should enable learners to practice using language in unrehearsed contexts, and help prepare them for an oral proficiency interview. However some students lack foundational skills in reading, pronunciation, and grammar, and these students will require reinforcement in these areas. The findings also support the hypothesis that students vary in terms of skill level and learning rate, and this slows progress of the class as a whole.

RELATED WORK

Previous work on Tactical Language (Johnson, 2010), VRP MIL (Johnson, 2015), and other related products (Johnson, 2012) incorporate virtual role-play activities in which learners can engage in simulated conversations with interactive characters. They also provide feedback on learner pronunciation. ALLEARN extends this work with activities that challenge learners to attain higher levels of spoken language proficiency, within an architecture that supports personalized, adaptive learning.

There are a variety of other digital tools available for language learning, such as CL-150 (Transparent Language, 2016) and Gloss (DLIFLC, 2016), but most do not incorporate speech recognition technology and this limits their effectiveness for practicing spoken language skills, or imposes burdens on instructors to evaluate learner speech recordings. For example, McGraw Hill Education's Conéctate product lets learners record themselves speaking Spanish phrases, but does not evaluate their speech. Learners get credit regardless of whether or not they are actually speaking Spanish. Examples of products that employ speech technology to address speaking skills include:

- Products that let learners practice scripted conversations, such as Rosetta Stone (Pellom, 2012). In this approach the software presents one or more prompts of what to say on the screen, the learner reads the prompt into the microphone, and the automated speech recognizer recognize it. This approach may be acceptable for rehearsing memorized dialogue, but it is not well suited for the improvised, unmemorized dialogue typical of more advanced spoken proficiency.
- Products that provide pronunciation feedback, such as Carnegie Speech (Eskanazi et al., 2007) and EnglishCentral. Although quality of pronunciation is one factor in assessing spoken proficiency, it is not the most important factor. A distinct foreign accent is acceptable as long as the learners can express themselves clearly in the target language.
- Products that employ speech recognition in specialized learning activities, such as DynEd (DynEd, 2016) and DuoLingo (Johnson, 2013). For example, DynEd uses speech recognition only in sentence unscramble exercises, where the learner assembles sentence fragments into a sentence and speaks it. Such activities may be useful but we believe much more can be done to provide learning activities that promote spoken proficiency and that fit well with instructor needs.

Although adaptive learning technology products are increasingly becoming available, most focus on mathematics and other STEM (science, technology, engineering, and mathematics) fields (EdSurge, 2016) as opposed to language and culture. Products such as Rosetta Stone incorporate adaptive sequencing algorithms, but are designed for use in self-study mode, as opposed to blended learning context. They are not designed to present material in a way that aligns well with a course curriculum, and that supports the particular topics that are being covered in class.

Adaptive training was first proposed as a problem of sequential optimization in training systems for the Air Force (Sebilske et al., 2009; Levchuk et al., 2012). Subsequent work in sequential optimization for Intelligent Tutoring Systems (ITS's) includes the RAPID tutor (Brunskill, 2010). In contrast to the current work, the RAPID tutor assumed full observability of student state (i.e., a Markov Decision Process instead of a Partially Observable

Markov Decision Process) and tracking of student skills as a binary (either the student possesses the skill or does not) rather than a graded continuum. Another planner was used on a Call for Fire (CFF) domain (Folsom-Kovarik, 2012), but this domain only used a handful of possible student states. This work builds upon our previous work with a factored representation (Horn et al,. 2010, Carlin et al., 2013, Carlin et al., 2016), which represents many state variables, each of which represents a graded continuum of skill levels, and which is partially observable. This previous work has been applied to security training, counter-piracy training, and adaptive analyst training software, but has never been applied to the language training or in a blended learning context, until the current work.

There is also a rich literature on Educational Data Mining (EDM). One approach is Bayesian Knowledge Tracing (BKT, Corbett & Anderson 1995). This approach models probability that a student knows a particular skill, a probability of guessing correctly anyway (if the skill is not acquired) or a small "slip" by getting a question wrong despite knowing the skill, and also the probability of learning the skill over time. This can formulate a Hidden Markov Model (HMM), but due to the tractability problems involved in reasoning about multiple skills, most previous work reasons about a small number of skills. Another EDM approach is Factor Analysis (e.g., Lan et al., 2014). This approach formulates a matrix of learners (columns) and questions (rows), and through dimensionality reduction, finds the fundamental skills. In contrast to BKT, which doesn't reason about a breadth of skills, typical Factor Analysis approaches don't reason about the evolution of skills over time.

ALLEARN SYSTEM ARCHITECTURE

Figure 1 shows the planned architecture for ALLEARN. We are developing it incrementally using our design science approach and plan to make adjustments as needed based on feedback from users and other stakeholders. The primary learner interface is the Alelo Enskill TM player, an HTML5-compliant cross-platform Web interface that incorporates speech recognition technology and supports a variety of learning activities. These activities enable learners to practice the spoken language skills and assess their performance. We are extending the range of Enskill activities to promote spoken language proficiency at the ILR 1+ to 2 level. We are providing 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 will support common learning content interoperability standards, such as the LTI (Learning Tools Interoperability) v2.0 standard (IMS, 2016). This will make it possible to integrate standards-compliant learning content from a variety of sources.

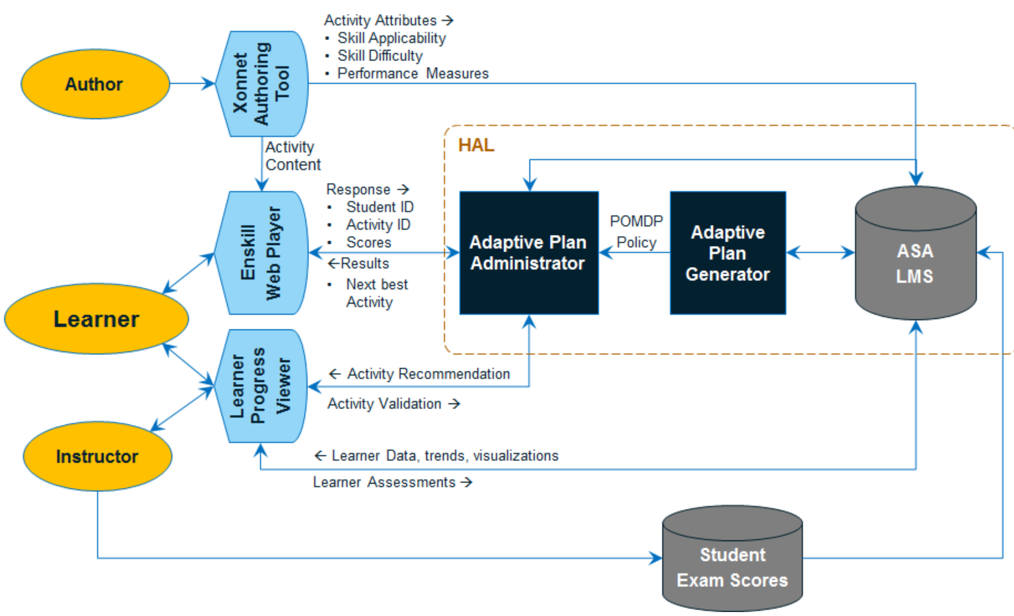

Figure 1. ALLEARN System Architecture

Instructional designers use the Xonnet authoring tool to specify content in a device-independent way, for delivery using the Enskill player. We plan to adapt this tool so that instructors will be able to use it as well. Content authored using other tools is run using the LTI launch mechanism.

At the heart of the system is an engine for adaptive recommendation of instructional content, labeled Higher Adaptive Learning (HAL) in Figure 1. The HAL capability includes an authoring-time component and a runtime component. The authoring-time component, called the Adaptive Plan Generator, analyzes the competencies to be trained and the instructional content, and generates a plan for selecting content to present to the learner, depending on the learner's level of mastery of the competencies. The plan takes the form of a POMDP (Partially Observable Markov Decision Process) policy, as described below. The Adaptive Plan Administrator executes this plan at runtime. It determines which instructional content to employ, based upon the learner's competency profile.

The Learner Progress Viewer is used to set student learning objectives and track progress. Instructors can compare their subjective assessments of student progress against the system-generated assessments. They can also review and edit the system's recommendations of activities for individual learners to work on. Students can set their own personal learning goals and track their progress toward achieving those goals. This encourages a learner-centered approach in which learners take greater responsibility for achieving their own learning goals.

LEARNING TECHNOLOGY DESIGN

Learning content will consist of microlearning and assessment modules, of at most 15 minutes in duration. Learning activities and assessments will center on spoken-language practice activities, together with other activities that reinforce supporting language skills. To practice language skills, learners engage in spoken interactions with on-screen computer characters. Some activities model individual conversational turns, while others model extended dialogs. To support varying levels of proficiency the dialogs will be playable at varying levels of difficulty (e.g., with or without hints), and incorporate non-player characters with nondeterministic behaviors. If the learner practices the role-play multiple times the play will be different each time. Learners thus get practice with improvised language use instead of relying on memorized phrases.

To accomplish the goal of reaching higher levels of language proficiency (ILR 2 or better), we break the general proficiency goal into more granular objectives that can be quantitatively measured and tracked. We employ the ACTFL proficiency guidelines (ACTFL, 2012) that map onto the ILR scale but which define finer gradations of proficiency in the range below ILR level 2 (novice low/mid/high, intermediate low/mid/high). Each level is further broken down into a set of communicative competencies with associated can-do statements, based on the NCSSFL-ACTFL inventory of can-do statements (ACTFL, 2015). Spoken-language practice activities are then created for the individual communicative competencies. If learners are able to demonstrate a variety of communicative competencies associated with a proficiency level at a sufficient level of performance, it provides strong evidence that the learner has achieved that level of proficiency. The Council of Europe's inventories of language skills at the Waystage and Threshold levels (comparable to ILR 1+ and 2, respectively) provide further detail regarding the language skills learners must demonstrate at each level (van Ek & Trim, 1991a & 1991b).

As multiple second language acquisition researchers have noted, spoken second language proficiency actually has three dimensions: fluency, accuracy, and complexity (Housen & Kuiken, 2009). There are tradeoffs among the three; as learners attempt to produce more complex language their fluency and accuracy tends to suffer, and vice versa. ALLEARN spoken-language activities automatically generate scores of complexity, accuracy, and fluency of the learner's input, along with completion scores that indicate how much the learner was able to complete without reliance on hints. This provides objective assessments of progress toward mastery of the target competencies.

Communicative competencies have enabling skills that support them. Enabling skills include basic foundational skills (e.g., pronunciation skills, reading skills, knowledge of grammatical concepts), knowledge of vocabulary, and enabling communicative competencies. These in turn will be linked to practice modules focusing on those skills. This makes possible a variety of possible paths through the learning materials. Some learners can spend more time practicing the enabling skills while others focus on practicing the communicative competencies in context.

For example, many learners in intensive courses struggle with vocabulary memorization. They cram vocabulary using flashcard memorization programs, only to forget it later. In ALLEARN we can provide learners with a combination of spoken-language activities that involve using vocabulary in context and memorization activities that focus on the enabling vocabulary knowledge. We can then determine which sequencing of lessons achieves the best long-term learning outcomes. It may be that it is better for learners to try using vocabulary in context first, before they attempt to memorize it, so that they build better associations in memory and are able to retain it.

ADAPTIVE LEARNING STRATEGY

Within Figure 1, the HAL Component consists of the Adaptive Plan Generator (during instructional design) and an Adaptive Plan Administrator (during execution). Figure 2 describes these components in further detail. These two components of the ALLEARN system make use of an underlying sequential optimization model, called a Partially Observable Markov Decision Process (POMDP). Figure 2 shows the components of a POMDP model, and how a solution to the model produces a set of rules called a Policy that is used in the Adaptive Plan Administrator to interact with the student.

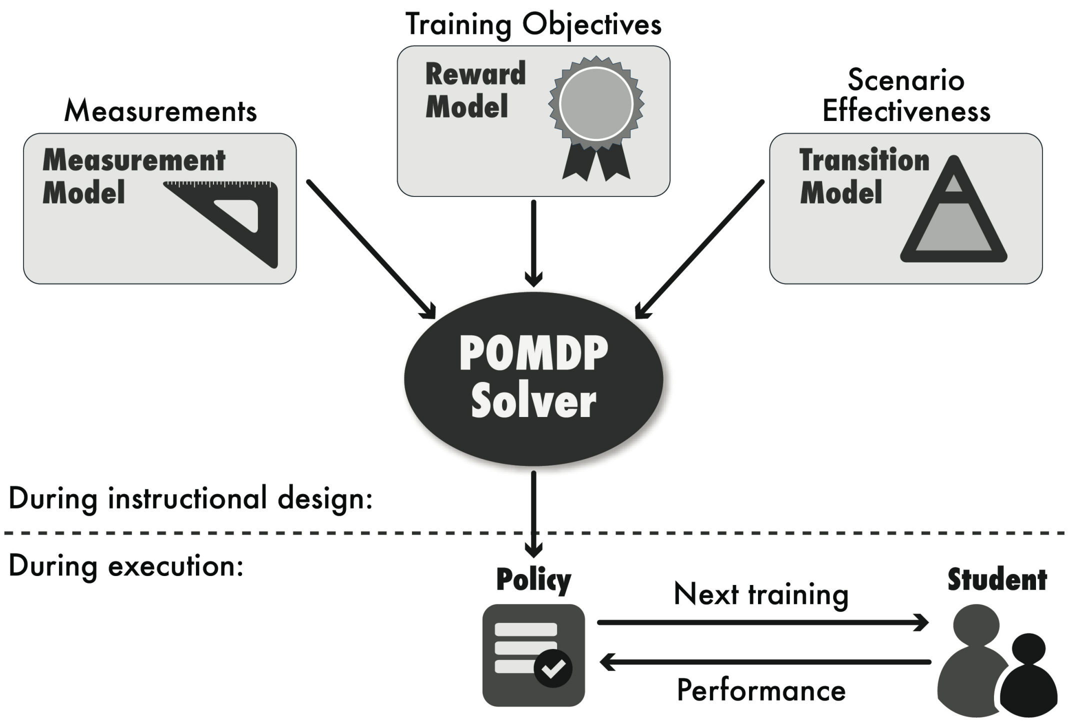

Figure 2. Adaptive Plan Generator.

Although POMDP is a general model, we configure its parameters (the transition model, reward model, and measurement model), to represent the language learning process. Part of the parameterized model includes instructor decisions, and in the model, student-learning outcomes depend the decisions made by the instructors, as well as the student's own personal learning state, as well as random chance. After creating the model, we then "solve" it, by generating a set of rules for the recommended instructor actions for any student profile. This set of rules is extremely large; for any history of student performance, we create a rule that recommends learning materials based on that student history. One can think of a rule as a software module that inputs student history, and outputs a recommendation for learning materials. One can think of the Adaptive Plan Generator as creating these rules, and an Adaptive Plan Administrator as executing the rule. An example of a rule is:

- Input to Rule: Student 111 took Lesson 1, followed by a 10-question quiz, 7 questions were correct but questions 3,5, and 10 were incorrect. Then the student took Lesson 1A followed by a 9/10 on the next quiz, with Question 7 incorrect.
- Output from Rule: The system recommends that the student take Lesson 2C next.

One can infer that the number of possible inputs to these rules is very large (even for software)! Mathematically, it is the number of permutations of student learning histories. Markov models (Markov, 1906) address this scalability problem by tracking student state. We model a student as existing in a *state*. A state is a set of possible student levels, we model student state as one of: Untrained, Novice, Partially Trained, Mostly Trained, Fully Trained. Our model extends state to address a variety of factors. That is, for each student list a set of skills (e.g., speaking, listening, reading, writing, decomposed further into "speaking days of week", etc.) and a level (e.g., Novice) associated with the student for each of the skills. Thus, in ALLEARN, an example of student state is:

- Speaking: <Days of week=Trained, Months of Year=Mostly Trained, Conjugation= Partially Trained>
- Listening: <Days of week=Trained, Months of Year=Mostly Trained, Conjugation=Partially Trained>
- Reading: <Days of week=Trained, Months of Year=Mostly Trained, Conjugation=Mostly Trained>

The mathematical purpose of representing student state is so that ALLEARN software can reason about student states, and not about the very large number of possible student histories. When a new event occurs in training (a new score on a quiz, notification that the student took a new lesson), student state is updated. Thus, the "Input to Rule" above, rather than being a student history, is a student state. Because measures are imperfect and thus student state is uncertain, our model tracks a belief distribution (i.e. a probability of) the student's state. Thus we revise the above to read:

- <u>Input to Rule</u>: 20% chance that student is Fully Trained on days of week, 50% chance Mostly Trained, 30% Partially Trained.
- <u>Output from Rule</u>: The system recommends that the student take Lesson 2C next.

One can observe that the rules above did not need to input the full set of factors in creating the rule (e.g., Months of Year was not included). This is one of the unique features of our approach, that our computational approach keeps the model tractable by only creating rules for a small set of factors at a time. To accomplish this, housekeeping is necessary to assure that rules do not conflict, and that rules cover every state. We do not document the housekeeping in this paper, but note that we follow the approach of (Guestrin, 2003).

Having defined state, we are now in a position to describe our adaptive training model more completely. Our model is shown in Figure 2. It consists of three components, all of which relate to the description of state above. Querying instructors, via an interface, develops the Measurement Model, Reward Model, and Transition Model. A POMDP solver inputs these models and outputs a Policy. This Policy is executive by the Adaptive Plan Administrator.

- <u>Reward Model</u>: Rule for assigning each state a value, indicating the desirability of having the student achieve a given state.
    - One common representation is to assign a value of "1" to Fully Trained, and a value of "0" to all other states. The training system will then derive that other states such as Partially Trained only have value insofar as they are necessary waypoints on the way to Fully Trained.
- <u>Measurement Model</u>: Rule for indicating how measures can be used to estimate student state. We use Item Response Theory (Lord, 1980) as a starting point. Our work extends IRT, by vectorizing IRT so that it represents multiple skills, while simultaneously implementing a Partial Credit Model (PCM) on this vectorized form.
- <u>Transition Model</u>: Rule for predicting how a learning material will affect student state. For example, the transition model could state that a video lesson on Days of the Week is 70% likely to improve a student from Untrained to Partially Trained, and 5% likely to improve that student to Fully Trained, and 25% likely not to improve the student at all.

The Adaptive Plan Generator is responsible for creating the above models. The Reward and Transition models are represented as probability functions. The Adaptive Plan Administrator uses those probability functions to recommend learning materials. It implements a function called State Update, which inputs the student's previous estimated state, and a new learning event. If the learning event was that the student took a lesson, the Adaptive Plan Administrator uses the Transition Model to update its assessment of the student state. If the learning event was that the student was measured (via a quiz, an instructor evaluation of speech, etc.), the Adaptive Plan Administrator uses the Measurement Model to update its assessment of the student state.

The Adaptive Plan Administrator then looks up, using its set of rules, the best learning material for a student in the assessed state. These rules are constructed by using the Reward Function, and reasoning. We follow method of MDP's (Puterman, 1994), while utilizing the factored structure (Guestrin, 2003). To summarize, our method is analagous to robot path planning, we assign value to each of the goal states at time t. Using this value, we assign value to states that are one step away from the goal states, at time t-1. Using these values, we assign value to states that are two steps away from the goal states, at time t-2, etc. The reasoning is somewhat complex, as in our model, state is never fully known (only imperfectly measured), and predictions are subject to variability. Our model reasons about these stochastic effects.

In summary, our adaptive learning algorithm selects the learning material (action) that will eventually lead to the highest reward (Fully Trained student state).

PLANS FOR SUBSEQUENT WORK

During the summer of 2016, we plan to review the ALLEARN design and user experience with learners and instructors and get their feedback. We then plan to continue development of the architecture, with the goal of undertaking a pilot evaluation at USAJFKSWCS in the spring of 2017.

The pilot evaluation will employ the following method. The pilot will cover the first two lessons of Module 2 of the curriculum, dealing with the exchange of personal information. These are Intermediate Low competencies that build on the Novice competencies in Module 1. At this point some students will be ready to make progress developing Intermediate-level competencies, while other students are still struggling with basic language skills. The current test at the end of Module 1, immediately prior to this unit, will provide evidence of these differences. We will provide an experimental group of learners with ALLEARN content covering the competencies addressed in the two lessons. Instructors will encourage the students to include ALLEARN in their study activities at their discretion. Because much of the learners' study activities are self-directed, they will be free to decide for themselves how to divide their time between ALLEARN activities and other self-study activities. ALLEARN will track the time the students spend, and estimate the students' degree of mastery of the competencies targeted in the lessons based upon their performance in the ALLEARN activities. Students will be tested at the end of the module. For comparison purposes we will ask an active group to use digital learning tools that they normally use, without ALLEARN, and track the time that they spend. We predict that time spent in ALLEARN activities will correlate with improved test performance. We will interview the instructors and students afterwards to get their perspective on the added value that ALLEARN provides.

After the system has been validated in this limited pilot trial we will then integrate it throughout the USAJFKSWCS course. At this point we should expect to see improved student progress overall, with greater numbers of learners achieving an ILR spoken proficiency score of 2 or better on an OPI.

Once the digital activities and learning model is fully developed we plan to adapt the Xonnet authoring tool to support authoring by instructors, not just trained instructional designers. Our goal is to provide an interface that is very easy to use and does not require technical expertise. A separate project named VRP Builder, funded by the National Science Foundation, is funding the development of natural language processing and authoring technology to enable authors, instructors, and ultimately students to create their own role-play dialogs from examples (Johnson & Koffler, 2016). This will make possible broad adoption of the approach.

CONCLUSIONS

The ALLEARN project has identified likely barriers to rapid acquisition and sustainment of spoken language skills, through detailed analysis of the needs of instructors and language learners. There is a lack of digital resources for practicing and assessing spoken language skills. There is a need for learning activities that can adapt to the individual learner. Technology-based resources are needed that integrate well with classroom activities, in a blended learning paradigm. We are iteratively developing a technical solution to address these needs, using a design-research methodology. 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 the ALLEARN into a model for adaptive language learning, both for civilian education and military training. As major 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.

ACKNOWLEDGEMENTS

The authors wish to thank everyone on the ALLEARN team, especially Kent Halverson, Brenda Lindsay, 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, Damon Regan, 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 the Defense Health Agency, Defense Language and National Security Education Office, Special Operations Forces Language Office, Office of Naval Research, or the Department of Defense.

REFERENCES

ACTFL (2012). *ACTFL Proficiency Guidelines 2012*. Alexandria, VA: American Council on the Teaching of Foreign Languages.

ACTFL (2015). *NCSSFL-ACTFL Can-Do Statements: Performance Indicators for Language Learners*. Alexandria, VA: American Council on the Teaching of Foreign Languages.

Brown, A.L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences,* 2(2), 141-178.

Brunskill, E., & Russell, S. (2010). RAPID: A reachable anytime planner for imprecisely-sensed domains. Conference for Uncertainty in Artificial Intelligence. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI).*

Carlin, A., Dumond, D., Freeman, J., & Dean, C. (2013, July). 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.

Chief of Naval Operations (2008). *U.S. Navy Language, Regional Expertise, and Cultural Awareness Strategy.* Washington, DC: Chief of Naval Operations.

Collins, Allan. (1992). Toward a design science of education. In Eileen Scanlon & Tim O'Shea (eds.) *New Directions in Educational Technology.* Volume 96 of the series NATO ASI Series pp. 15-22.

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction, 4(4), 253-278.

DLIFLC (2016). Gloss. Retrieved May 19, 2016 from https://gloss.dliflc.edu/.

Department of Defense (2011). Strategic plan for language skills, regional expertise, and cultural capabilities: 2011-2016. Retrieved May. 15, 2016 from http://prhome.defense.gov/Portals/52/Documents/RFM/Readiness/DLNSEO/files/STRAT%20PLAN.pdf

Donnelly, J. (2015). SOF Language Program Review. Presentation at LREC Industry Day, Tampa, FL, 17 December 2015.

DynEd (2016). Speech Recognition Basics. Retrieved May 19, 2016 from https://msu.edu/course/esl/094/Dyned%20V2.7/DynEd/doc/srguidee.pdf.

EdSurge (2016). *Decoding Adaptive.* London: Pearson.

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.

Folsom-Kovarik, J. T. (2012). *Leveraging Help Requests in POMDP Intelligent Tutoring Systems* (Doctoral dissertation, University of Central Florida Orlando, Florida).

Freeman, H. & Cohn, J. (2015). S&T for Blended Adaptive Language & Culture Training Symposium. *HPT&B Newsletter* 3, p. 21.
Guestrin, Carlos, Daphne Koller, Ronald Parr, and Shobha Venkataraman. "Efficient Solution Algorithms for Factored MDPs." Journal of Artificial Intelligence Research 19 (2003): 399-468. JAIR. Journal of Artificial Intelligence Research, 3 Oct. 2003.
Horn, Z., Freeman, J., & Carlin, A. (2011). Adaptive Device for Adaptive Performance Training (ADAPT). Final Report presented to Office of Naval Research. Aptima.
IMS (2016). Learning Tools Interoperability. IMS Global Learning Consortium. Retrieved May 27, 2016 from https://www.imsglobal.org/activity/learning-tools-interoperability
Housen, A. & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics* 30(4), 461-473.
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. *Int. J. of Artificial Intelligence in Ed*. 20(2), 175-195.
Johnson, W.L. (2012). Error detection for teaching communicative competence. Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training.
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. & Koffler, R. (2016). Alelo and USC Institute for Creative Technologies Announce Partnership to Co-Develop Disruptive Learning Technologies. Retrieved May 19, 2016 from http://www.prweb.com/releases/2016/01/prweb13162572.htm.
Lan, A. S., Waters, A. E., Studer, C., & Baraniuk, R. G. (2014). Sparse factor analysis for learning and content analytics. *The Journal of Machine Learning Research*, *15*(1), 1959-2008.
Lord, F. M. (1980). Applications of item response theory to practical testing problems. Routledge.
Markov, A. (1906). Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga. *Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete*, 2-ya seriya. pp. 135–156
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.
Shebilske, W., Gildea, K., Freeman, J., Levchuk, G. (2009). Optimizing Instructional Strategies: A Benchmarked Experiential System for Training (BEST). Theoretical Issues in Ergonomic Science. Special Issue on Optimizing Virtual Training Systems, v10 (3) May 2009 , pages 267 - 278.
Shearer, S.R. (2013). *Modeling Second Language Change using Skill Retention Theory.* PhD Dissertation, Naval Postgraduate School.
SWA Consulting (2010). *US Army Special Forces Language Proficiency Needs Assessment.* Tech report no. 2010010623. Raleigh, NC: SWA Consulting.
Tokunaga, M. (2010). Metalinguistic knowledge of low-proficiency university EFL learners. In A.M. Stroke (Ed.), *JALT 2009 Conference Proceedings.* Tokyo: JALT.
Transparent Language (2016). CL-150 Technology Matrix for Critical Languages. Retrieved May 19, 2016 from http://www.transparent.com/government/about-cl-150.html.
Van Ek, J.A & Trim, J.L.M (1991a). *Waystage 1990*. Cambridge, UK: Cambridge University Press.
Van Ek, J.A & Trim, J.L.M (1991b). *Threshold 1990*. Cambridge, UK: Cambridge University Press.
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.