Tactical Language and Culture Training Systems:

Using AI to Teach Foreign Languages and Cultures

W. Lewis Johnson and Andre Valente

■ *The Tactical Language and Culture Training System (TLCTS) helps people quickly acquire communicative skills in foreign languages and cultures. More than 40,000 learners worldwide have used TLCTS courses. TLCTS utilizes artificial intelligence technologies during the authoring process and at run time to process learner speech, engage in dialogue, and evaluate and assess learner performance. This paper describes the architecture of TLCTS and the artificial intelligence technologies that it employs and presents results from multiple evaluation studies that demonstrate the benefits of learning foreign language and culture using this approach.*

*T*he Tactical Language and Culture Training System (TLCTS) helps people quickly acquire functional skills in foreign languages and cultures. It includes interactive lessons that focus on particular communicative skills and interactive games that apply those skills. Heavy emphasis is placed on spoken communication: learners must learn to speak the foreign language to complete the lessons and play the games. It focuses on the language and cultural skills needed to accomplish particular types of tasks and gives learners rich, realistic opportunities to practice achieving those tasks.

Several TLCTS courses have been developed so far. Tactical Iraqi, Tactical Pashto, and Tactical French are in widespread use by U.S. marines and soldiers, and increasingly by military service members in other countries. Additional courses are being developed for use by business executives, workers for nongovernmental organizations, and high school and college students. While precise numbers are impossible to obtain (we do not control copies made by the U.S. government), over 40,000 and as many as 60,000 people have trained so far with TLCTS courses. More than 1000 people download copies of TLCTS courses each month, either for their own use or to set up computer language labs and redistribute copies to students. Just one training site, the military advisor training center at Fort Riley, Kansas, trains approximately 10,000 people annually.

Artificial intelligence technologies play multiple essential functions in TLCTS. Speech is the primary input modality, so automated speech recognition tailored to foreign language learners is essential. TLCTS courses are populated with "virtual humans" that engage in dialogue with learners. AI techniques are used to model the decision processes of the virtual humans and to support the generation of their behavior. This makes it possible to give learners extensive conversational practice. Learner modeling software continually monitors each learner's application of communication skills to estimate the learner's
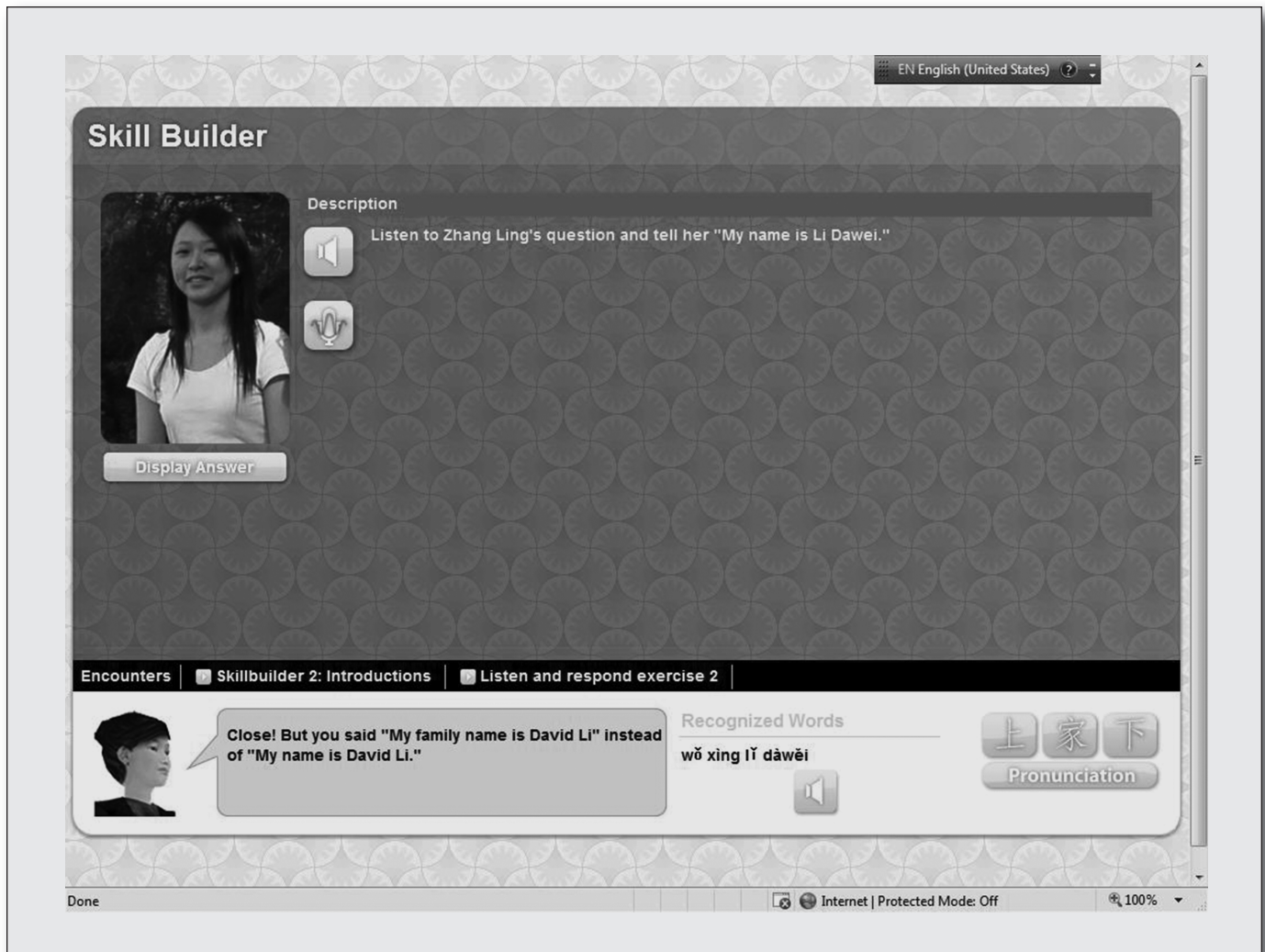
Figure 1. Example Skill Builder Exercise.

level of mastery of these skills. This helps instructors and training supervisors to monitor learners' progress and enables the software to guide learners to where they need to focus their training effort. Artificial intelligence is also integrated into the systems' content authoring tools, assisting content authors in the creation and validation of instructional content.

System Overview

TLCTS courses are currently delivered on personal computers, equipped with headset microphones. Each course contains a set of interactive Skill Builder lessons, focusing on particular communicative skills. Figure 1 illustrates an exercise page from the Skill Builder of *Encounters: Chinese Language and Culture,* a college-level Chinese course being developed in collaboration with Yale University and Chinese International Publishing Group. The figure is an example of a minidialogue exercise where the learner practices a conversational turn in the target language. Here the learner must think of an appropriate way to say his name. His response, "Wǒ xìng Lǐ Dàwěi" (My family name is Li Dawei) exhibits a common mistake made by beginning language learners: to confuse the semantically similar words "xìng" (family name is) and "jiào" (name is). A correct response would be "Wǒ jiào Lǐ Dàwěi." The virtual tutor (bottom left) gave appropriate corrective feedback. The spoken conversational system recognized the utterance (bottom center), and was able to recognize the error in the utterance, so that the virtual tutor can respond accordingly. Note that in exercises there is no single correct answer: any utterance that correctly conveys the intended meaning will be accepted. Depending on the type of exercise, the system can give feedback on pronunciation, morphological and grammatical forms, cultural pragmatics, or word choice, as in this example.

Figure 2. Example Mission Game Dialogue.

Games play an essential role in TLCTS courses, providing essential practice opportunities. Each course incorporates a scenario-based Mission Game, where learners play a character in a three-dimensional virtual world that simulates the target culture. Figure 2 shows a screenshot from the Mission Game in Tactical Dari, currently being used by U.S. military service members to learn the Dari language and Afghan culture in preparation for deployment to Afghanistan. Here the player (left) is engaged in a meeting with a village leader (the malek) to discuss reconstruction plans for a local school. A transcript of the conversation to this point is shown in the top center of the screen.

The learner speaks to the malek by first clicking the microphone icon (top right) and then speaking into a headset microphone. He or she is free to discuss a range of topics (far top left) and express each topic in a range of ways (top left, below). Although this screen shot is displaying help menus that make these choices explicit, in ordinary use

these choices are hidden, and the learner is encouraged to engage in free spoken conversation in the foreign language in the topic at hand. This contrasts radically with typical uses of speech recognition in other language-learning systems, which either do not support spoken dialogue at all or present learners fixed sets of choices to read off of the screen.

The Alelo Architecture

The architecture underlying TLCTS (figure 3) supports a set of complementary training products built from a common set of content specifications. All content in TLCTS courses is specified in Extensible Markup Language (XML) and stored in a web-compatible content repository. A web portal named Kona[1] provides access to the repository and supports collaborative editing. Kona supports a collection of web-based authoring tools that enable authors to edit and modify content specifi-
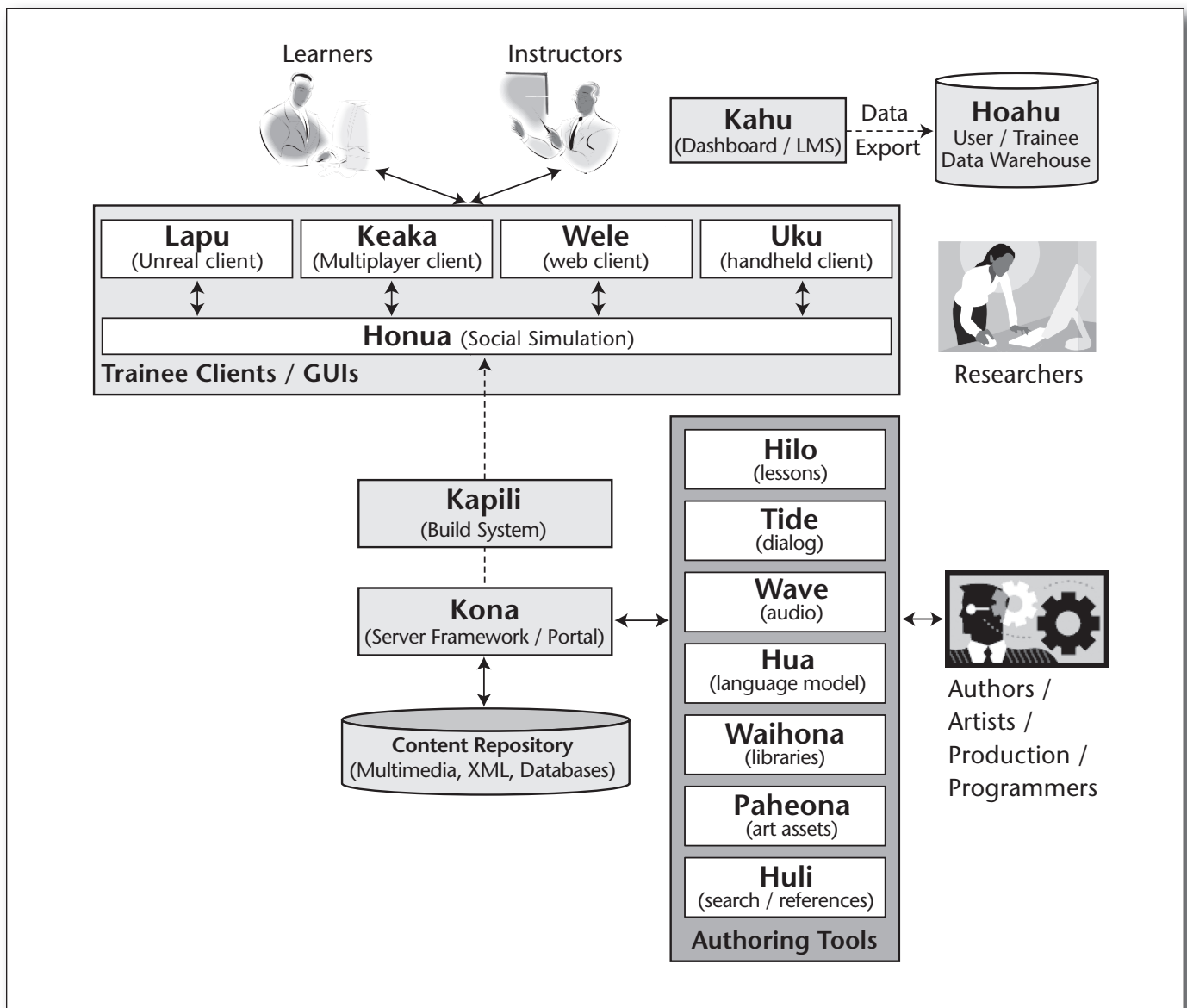
Figure 3. Overall Alelo Architecture.

cations using their web browsers. The authoring tools are integrated in the portal, but each tool applies to a different subset of the content. One tool (Hilo) edits lesson content, while another (Tide) edits dialogue specifications within lessons and game episodes. Production staff edit and manage speech recordings using the Wave audio tool. The Hua editor is responsible for managing the language model describing the target language: words and phrases, their spellings, phonetic transcriptions, reference recordings, and translation. Nonverbal gestures and character models are defined using Waihona and Paheona, respectively.

Each project is represented in Kona as a "book," containing a series of "chapters" (each specifying an individual lesson or game scene). Project mem-

bers can check out individual chapters (that is, lock them for editing), update them using web-based authoring tools, and then check them back in. An underlying version-control framework (Subversion) prevents simultaneous edits and allows rollback to earlier versions if necessary. Our goals have been to make the authoring cheaper, faster, better, and easier and to eliminate the need for engineers and researchers to be involved in authoring and production.

The architecture supports several delivery platforms. Our Lapu client is the first client that we developed, and is still the most widely used. Lapu runs on a Windows PC equipped with a headset microphone. Lapu is built on top of Epic Games' Unreal Engine 2.5, which handles scene rendering
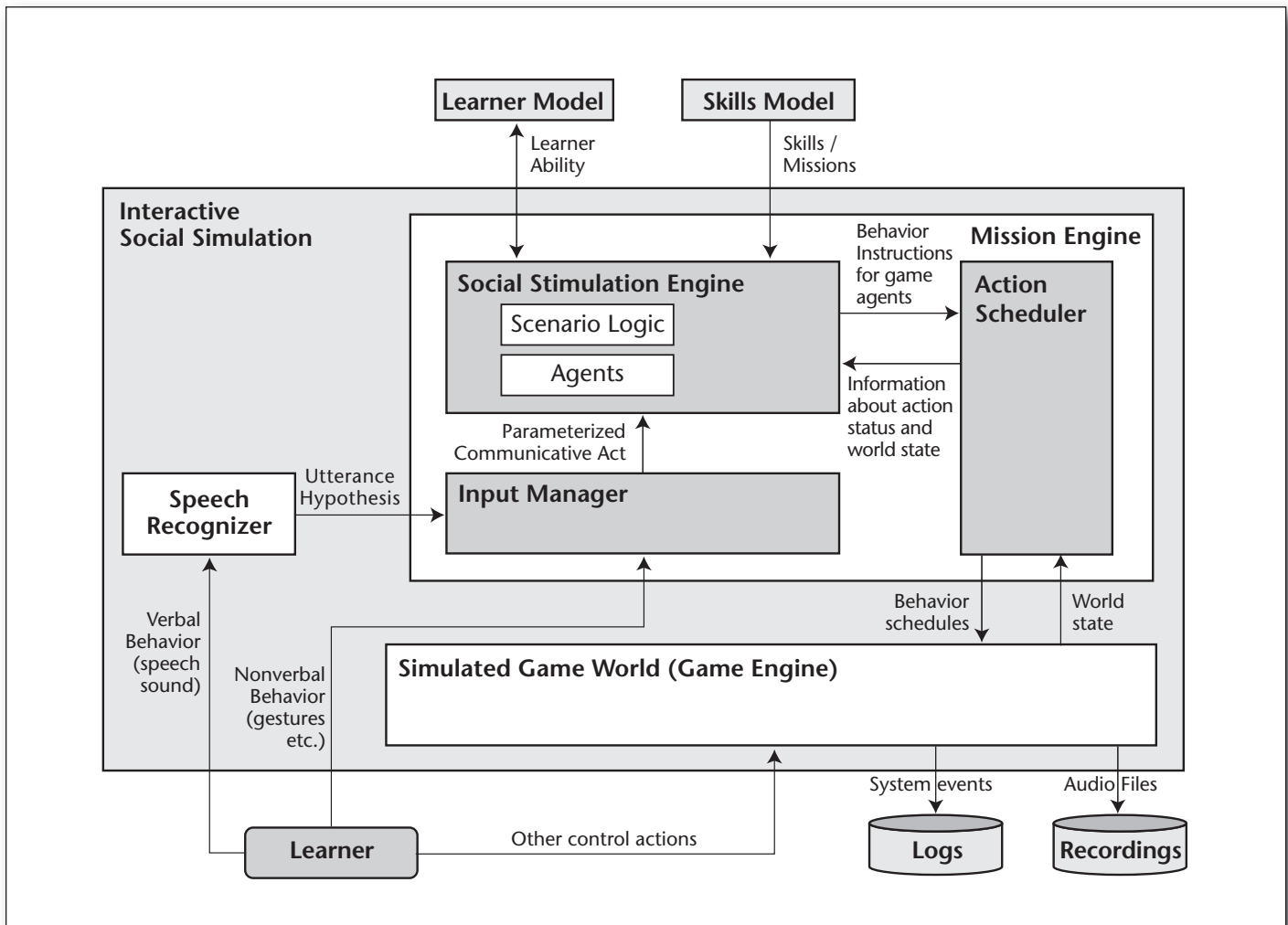
Figure 4. Lapu Client Simulation Architecture.

and provides a set of user interface classes. Figure 4 shows the Lapu architecture in further detail, focusing on the support for social simulations (dialogues with animated characters). The learner communicates with characters in the game, using voice and gestures selectable from a menu. The input manager interprets the speech and gestural input as a communicative act (that is, speech acts augmented with gestures). The social simulation engine (Honua) determines how the environment and the characters in it respond to the learner's actions. The character actions are realized by the action scheduler, which issues animation commands to the unreal engine. As the learner interacts with the system, the learner model is updated to provide a current estimate of learner skill mastery. Interaction logs and learner speech recordings are saved for subsequent analysis. These components are implemented primarily in Python, with some supporting routines in C++ and UnrealScript (the scripting language for the unreal engine).

Reliance on UnrealScript is minimized to facilitate porting to other game engines.

We have recently developed a multiplayer extension (Keaka) that enables multiple learners to interact with the virtual world at the same time. Each learner interacts with one character at a time, can overhear the conversations of other learners, and can also talk to other learners through voice over IP. Keaka is being used to support mission rehearsal exercises, in which teams of learners must speak with nonplayer characters in order to gather information necessary to complete their mission.

A new web-based client named Wele is also being increasingly used to deliver content. Wele currently supports the Skill Builder, and we are extending it to include Mission Game implementations. Wele is accessed through a web browser and is implemented in the Adobe Flex[2] Internet application framework. Wele runs the speech recognizer on the client computer as a plug-in. Inter-

active dialogues and animations are realized as Flash files inserted into a web page and are controlled at run time using finite state machines.

Handheld device implementations (Uku) allow trainees to continue their training when they do not have access to a PC. One version of Uku currently under beta test delivers TLCTS Skill Builder content on iPods. Media assets (instructional texts, recordings, images, and videos) are extracted from the repository, converted into iPod-compatible formats, and organized using the iPod Notes feature[3]. Example dialogues between animated characters are converted into videos. We developed a prototype for the Sony PlayStation Portable, and we are evaluating other handheld platforms. All clients except for the iPod provide interactive exercises. In cases where we are able to port the speech recognizer to the handheld device, the client implementation utilizes the speech recognizer; for platforms that cannot run the speech recognizer effectively, speech-recognition input is replaced with alternative methods, such as selection from menus of possible utterances.

A lightweight learning management system called Kahu communicates with the client computers over a local network. Our users frequently organize their computers in learning labs using ad hoc local networks (using Windows shares), then move their computers around and reorganize them in different configurations. Kahu provides and manages repositories for learner profiles and supports easy reconfiguration and disconnected operation of these labs. The system provides mechanisms for an instructor or training manager to create and configure users and groups of users and produce progress reports. We also use it to help retrieve learner profiles, system logs, and recordings from learning labs and store them in a data warehouse, where they may be retrieved for further analysis. The recordings are used to retrain and improve the speech recognizer, while the logs and learner profiles are used in research to understand how the users employ the learning environment and to suggest areas for further improvement of the software.

Uses of AI Technology

Artificial intelligence technologies are employed extensively in the Alelo run-time environments and authoring tools. The following is a summary of the roles that artificial intelligence currently plays, starting with the run-time environment. However, in practice, authoring concerns and run-time concerns are inextricably linked; run-time processing methods cannot be employed if they place an unacceptable burden on authors to create content for them.

Authoring

Our authoring tools are designed to help authors create the rich content representations required by our AI-based system, and perform AI-based processing themselves. For example Hilo, our lesson authoring tool, supports rich utterance representations for speech and natural language processing. Utterances are modeled in different channels—the native orthography of the foreign language, an "ez-read" transliteration that is intended as a phonetic transcription in Roman characters, a phoneme sequence used by the speech recognizer, and an English translation. Each utterance is also linked to the language model library so it can be centrally managed.

For some languages we augment the authoring tools with tools that translate between channels. Chinese and French provide two contrasting examples of how this is done. In Chinese, authors specify the written form in traditional Chinese characters and specify the ez-read form in Pinyin. Then the run-time system translates automatically from traditional characters to simplified characters as needed and translates from Pinyin to phoneme sequence for speech recognition purposes. For French we provide authors a tool that generates phoneme sequences from written French and then generates an ez-read view from the phoneme sequences. French spelling is highly irregular, so the phoneme sequence generator is not fully automatic. Instead the phoneme sequence generator proposes alternative pronunciations, based on a lexicon of French words, and lets the author choose the correct phoneme sequence.

Our Tide authoring tool provides several editing functions and analysis tools to specify interactive dialogues. Authors specify the utterances that may arise in the course of a conversation, their meaning in terms of acts, and the possible interactions between acts in the conversation. Tide supports several dialogue models, from simple scripts to branched stories to agent-centered models. It provides graphical tools with a custom vocabulary of abstractions to make interaction authoring as easy as possible. From the specifications, it automatically generates program code, which can be incorporated into the TLCTS clients. Finally, Tide provides a model-checking tool for testing dialogue modeling codes against the specification. This makes it possible for authors to produce a substantial portion of the dialogue modeling work themselves.

Our Hua tool manages the language model, which contains the library of words and utterances that are taught in the course. It interoperates with Hilo and Tide as needed so that as authors edit content they can maintain a consistent language model throughout. It also provides for different roles: content developers can add utterances quick-

Content Metric	Tactical Iraqi 3.1	Tactical Iraqi 4.0	Tactical French 1.0	Tactical Dari 1.0
Lessons	35	52	40	51
Lesson pages	891	2027	1920	2139
Words	1072	2214	1820	1199
Example dialogues	42	85	67	44
Active dialogues	13	29	36	31

Table 1. Content Size of TLCTS Courses.

ly but that content is later verified and corrected by more senior language experts playing the role of "librarian." This helps to eliminate errors and reduces the role of natural language processing specialists in creating and maintaining linguistic resources.

These authoring tools have enabled us to increase the quality and quantity of authored content and have made it possible to maintain multiple versions of content for different classes of learners and different platforms. Table 1 shows some of the productivity improvements that these tools are yielding. It compares Tactical Iraqi 3.1, developed in 2006, Tactical Iraqi 4.0, developed in 2007, Tactical French 1.0, also developed in 2007, and Tactical Dari 1.0, developed in 2008. The authoring tools doubled the number of lesson pages, vocabulary words, dialogues, and Mission Game scenes in Tactical Iraqi 4.0 vis-à-vis 3.1. Development of Tactical French 1.0 did not start until late 2006, yet it has about the same amount of Skill Builder material as Tactical Iraqi. Tactical Dari 1.0 was developed in less time and on a considerably smaller budget than Tactical French, and yet it is comparable in size (vocabulary is smaller because it is a much more complex language).

Speech Recognition

TLCTS is particularly ambitious in its reliance on speech-recognition technology. Recognition of learner speech is particularly demanding and challenging. Beginning language learners sometimes mispronounce the foreign language very badly and blame the software if it is unable to recognize their speech. The speech recognizer also needs to be very discriminating at times and accurately detect speech errors to provide feedback to learners. Typical speech recognizers, in contrast, are designed to disregard speech errors and focus on decoding speech into text. Our approach must apply equally to common languages such as French and to less commonly taught languages such as Pashto or Cherokee, for which few off-the-shelf speech-

recognition resources exist. For these reasons we rejected prepackaged speech-recognition solutions and opted to develop our own tailored speech-recognition models.

We employ hidden Markov acoustic models developed using the hidden Markov model toolkit (HTK)[4] and delivered using the Julius[5] open source speech-recognition toolkit. Models are developed iteratively in a bootstrapping process. We construct an initial acoustic model from a combination of licensed corpora, an initial corpus of learner speech, and sometimes corpus data from similar languages and integrate it into the first beta version of the TLCTS course. We then recruit beta testers to start training with the software. It records samples of their speech, which we use to retrain the speech recognizer. Each successive version of a TLCTS speech recognizer has improved performance over the previous versions because it is trained on speech data collected from the previous versions.

The speech recordings must be annotated to indicate the phoneme boundaries. Speech annotation is time-consuming but has a critical effect on the quality of the resulting acoustic model. Corpora licensed from other sources often have annotation errors and need to be reannotated.

Each language model includes a grammar-based model built from phrases extracted from the authored content and a default "garbage model" constructed with a fixed number of acoustic classes, states, and probability mixtures. If the learner speaks an utterance that is in the recognition grammar, the grammar-based model will usually recognize it with higher confidence than the garbage model does, and if the utterance is out of grammar the garbage model will recognize it with higher confidence. The garbage model rejects many utterances that the learners know are incorrect, and forces learners to speak the language with at least a minimum level of accuracy. By adjusting the properties of the garbage model we can adjust the recognizer's tolerance of mispronounced speech.

Language learners normally produce several types of errors in their speech, including pragmatic errors (for example, failure to adhere to cultural norms in discourse), semantic errors (for example, word confusions), syntactic errors, morphological errors, and pronunciation errors. Where possible we use the speech recognizer to explicitly detect these errors, using language models that are designed to detect language errors. Pronunciation error detection is handled as a special case. Earlier versions of TLCTS attempted to detect pronunciation errors on a continual basis in the Skill Builder (Mote et al. 2004). However, evaluations identified problems with this approach: it was difficult to detect pronunciation errors reliably in continuous

speech (leading to user frustration), and the continual feedback tended to cause learners to focus on pronunciation to the exclusion of other language skills. We have since adopted a different approach where the system does not report specific pronunciation errors in most situations, but instead provides a number of focused exercises in which learners practice particular speech sounds they have difficulty with.

Speech-recognition performance depends in part on the authored content, and so we have developed authoring guidelines and tools that enable content authors to create learning materials suitable for speech recognition without requiring detailed knowledge of the workings of speech-recognition technology. When authors write dialogues and exercises, they specify alternative ways of phrasing these utterances, some of which are well formed and some of which may illustrate common learner errors. These are used to create recognition grammars that recognize learner speech as accurately as possible. Authors can mark phrases as "ASR-hard," meaning that the automated speech recognizer (ASR) should not be applied to them and they should be excluded from the speech-recognition grammar. Since speech-recognition accuracy tends to be better on longer phrases, individual words (which sometimes appear on lesson pages) are sometimes excluded in this way.

We have used learner data to evaluate speech-recognition performance for the current languages and obtained recognition rates of 95 percent or above for in-grammar utterances. However the training data is skewed towards male users and beginner users (reflecting our current user population), which makes the performance degrade for female and advanced users. Learners with a variety of accents, including regional American accents and foreign accents, have used the system and have obtained adequate recognition rates.

Despite these positive results, users have raised issues about the speech-recognition function. For example, some users complained that the recognizer did not recognize what they thought were valid utterances. These problems can be caused by several reasons. First, some acoustic models (for example, for French) were built using proportionally more native data (than nonnative) and as such are less forgiving of mispronunciations. Other users complain that they found the ASR too lenient. We have responded to these problems in part by allowing the learner to individually adjust the garbage model (see above) to make it appear more or less lenient of mistakes. A second reason for these problems is that the ASR uses grammars that change for each specific part of the system. These grammars are compiled from existing content, and thus reflect whatever subset of the language we are teaching. Therefore, a native or fluent

speaker usually manages to produce an utterance that is not recognized because it is outside of the system's grammar. We have been working on techniques to extend these grammars while maintaining recognition rates (Meron, Valente, and Johnson 2007). Finally, many complaints assumed unrealistic expectations as to the type of noise the system could sustain while keeping recognition rates. A user sent us some system recordings with a loud TV program in the background and still complained that the recognizer was not working well. In response, we developed a speech quality assessment module that gives the learner feedback about possible indications of poor voice recording—levels too low or too high, excessive periods of silence, noise, or indications that the beginning or end of the recording was clipped. This is displayed first as the background color on a small waveform display—red for poor recording quality, green for good, yellow for marginal. If a user clicks that display, a larger window appears that compares the learner's speech with the native speaker's.

Modeling Dialogue

One key challenge for the dialogues in TLCTS is to manage the variability of the spoken inputs a learner may produce. On one hand, we wish to attain high variability, meaning that the system should recognize, properly interpret, and react to a large subset of the language. On the other hand, we wish to author interactions that train specific language and culture skills, not unconstrained dialogue. Furthermore, increasing the size of the language subset makes speech recognition and interpretation much harder. Our strategy to balance those needs is to manage variability in two layers. In the speech layer, we constrain the speech to be recognized in such a way that we can limit the possible utterances and improve ASR performance as described above. In the act layer, we manage the possible acts the user can perform and the mapping from speech to acts.

We originally created linear scripts and branched scripts and annotated each utterance with an act name. We then manually translated the text scripts into hard-coded programs. This translation included the logic for plot progression and character control, as well as the mapping from user input (the text output by the ASR) to actions in the game world (in the form of a list of input-action pairs). However, we found out that scripts were too restrictive a medium in that they limited the act variability. We could improve the situation by manually programming more complex algorithms in software code, but that approach required a programmer to complete the scene (which was both expensive and nonscalable), and it was impossible to verify whether the code was consistent with the author's intent.

In our current methods, authors write specifications of dialogue interactions at the act level. To increase the variability of dialogue that the framework supports, we introduced utterance templates into the dialogue specifications. An utterance template is a combination of a more flexible grammar definition syntax (which allows increased variability in user inputs) with additional syntactic constructs used to specify semantic content allowing parameterized utterances (which increases variability in possible responses). The grammar is passed directly to the speech recognizer, and when an utterance is parsed it comes with semantic annotation that indicates information such as the type of action or action-object specified in the utterance. We have developed tools maintaining these utterance templates and have been working on mechanisms for managing libraries of utterance templates (Meron, Valente, and Johnson 2007).

Believable Virtual Humans

A central feature of our approach is to put the user in a social simulation with nonplayer characters (NPCs)—*virtual humans*. Our original virtual human behavior generation pipeline was relatively simple. An input from the player would be provided as an abstract act (say, *greet_respectfully*) to an agent implemented in PsychSim (Si, Marsella, and Pynadath 2005). The agent system would specify an act for its NPC to perform (for example, *inquire_identity*). An XML file would specify the utterance and any animations that needed to be performed to realize that act; sometimes also a small script in Python would be called to perform more complex physical actions (say, walk over a specific place). More recently, we have adopted a variant of the SAIBA framework (Vilhjalmsson and Marsella 2005), which separates intent planning (the choice of courses of action—to complain, cooperate, make a request, and so on—that are adequate to the situation at hand) from the production of believable physical behavior.

We originally used the PsychSim agent system to perform intent planning (Si, Marsella, and Pynadath 2005). While PsychSim is highly expressive, its agents are extremely hard to author and computationally complex. The modeling language is not well suited to represent dialogues or world knowledge. We then turned to a finite-state machine approach that authors can use to specify moderately complex scenarios and that has proved scalable to large numbers of dialogues (for example, Tactical French has more than 50). We developed additional idioms to help us author certain conversation patterns. For example, we use range-qualified transitions to implement a simple form of the "beat" concept as used in Façade (Mateas and Stern 2005). This helped us more easily specify bits of dialogue that can be used within certain phases of the story but do not directly change the story state beyond indirect effects such as increasing rapport. We are now designing a third generation of intent planning called LightPsych, which is an integrated framework where agents can operate at any of four levels—self (explicitly model beliefs, short- and long-term memory, and so on), culture (model and reason with politeness, and social norms), physical and social environment and dialogue (understand dialogue openings and closings, requests, turn-taking, negotiations, and the meaning of utterances). A key challenge for LightPsych is to make it easy to create and reuse agent-behavior specifications and handle the most common interactions between trainees and NPCs but also to give authors tools to add to and modify these specifications to handle the aspects that are unique to each scenario.

A critical issue is the representation of rich communicative acts. We started representing acts as strings that encoded some of the parameters needed by PsychSim agents. However, to accurately represent the richness of conversational speech, we need a richer representation for speech acts. We looked at frameworks such as FrameNet (Baker, Fillmore, and Lowe 1998) before creating our own ontology of communicative acts based on the typology proposed by Traum and Hinkelman (1992) and extended with new acts (for example, offering greetings, thanks, support, and so on) based on the work by Feng and Hovy (2006). We also added representations of conversational context, such as the characteristics and roles of interlocutors, and the state of the world in which the conversation takes place. Our new act representation format extends the functional markup language (FML) but extends it in a number of ways (Samtani, Valente, and Johnson 2008).

With respect to the generation of animated behavior, we have found that our initial hand-coded solution was optimally flexible but hampered our efforts to lower costs and speed development. Further, these scripting or motion capture approaches may work if the social context is fixed and known ahead of time but break down if the social environment is dynamic. Therefore, we started working on automating the production of behavior animation from intent and are exploring the use of representations such as the behavior markup language (BML).[6]

Learner Modeling

As learners use TLCTS courses, it is important to track whether the learners are making progress toward learning objectives. Evidence for progress can come from multiple sources: the trainees' performance on games and quizzes, their performance in dialogues, and their performance generally in the TLCTS games. TLCTS records the learners' per-

formance in quizzes and dialogues, but also attempts to estimate the learners' mastery of key skills, using a technique inspired by the model-tracing technique of Corbett and Anderson (1995). Authors annotate exercises, quiz items, and dialogue exchanges to indicate the skills that they employ. Skills are drawn from a taxonomy of enabling learning objectives, encompassing linguistic skills, cultural skills, and task skills. Each correct or incorrect use of a given skill provides probabilistic evidence of mastery of that skill. This evidence is uncertain because learners may guess an answer or slip and make an unconscious mistake, or the speech recognizer may misinterpret the learner's response. However, after a relatively short amount of training time the learner model is usually able to correctly identify the individual skills that the trainee has mastered.

This learner model provides learners, trainers, and researchers with a rich view of trainee progress. We plan to use it as a basis for automated provision of training guidance, to advise learners about where they should focus their efforts, and when to adjust the amount of scaffolding to increase or decrease the level of challenge of the game. We also plan to use it as a starting point for estimating learner progress toward achieving longer-term learning objectives such as general spoken language proficiency.

Application Development and Deployment

The TLCTS system developed over several years, starting as a research project at the University of Southern California (USC) Information Sciences Institute and later at Alelo. The original project at USC spanned about four years starting in 2003, and its continuation at Alelo started in 2005. Costs for the development of the underlying science and technology have been around $5 million to date; costs for the engineering work to make it deployable have been more than $1 million, and course development costs have gone down from the $800,000 range to $300,000–$600,000, depending upon the target language and culture. This was first funded by Defense Advanced Research Projects Agency (DARPA) and U.S. Special Operations Command (USSOCOM) and more recently by U.S. Marine Corps, Army, and Air Force and the Australian Defence Forces.

Throughout the project we have adopted an iterative development strategy. The goal is to produce new versions of the system every several months, deploy and evaluate these new versions, and use the evaluation results to inform the next spiral. We believe that research and development benefit from use in practice and that each real-life testing stage provides critical feedback. Indeed, we believe that a

good deal of the success of TLCTS has stemmed from acting upon the feedback from our users.

Some of the key development challenges have been related to delivery. The transition from a research prototype to a widely used application required significant engineering effort to develop good installers, integrate the components more closely, improve performance, add user interface polish, and so on. The early version of the architecture was intended as a platform for experimentation (for example, using agent-oriented integration mechanisms), but the architecture then had to be simplified and streamlined to be stable and fast enough for a deployed system. At several points we had to sacrifice flexibility for performance, stability, and ease of use. For example, the original implementation of the Skill Builder was built on top of a separate courseware platform (ToolBook), but users found it cumbersome to switch back and forth between separate Skill Builder and Mission Game applications. We therefore converted the Skill Builder to run on top of the game engine, which was unfortunately less flexible—for example, only later were we able to recover the ability to play videos in our lessons.

Just as we have restructured and improved the system architecture, we have progressively improved the curricula that TLCTS supports. This has often involved incremental revisions of substantial bodies of learning content, as well as the representations of that content. We have therefore invested significant effort to developing authoring and conversion tools (see the section on Authoring) to support this maintenance process. Our use of XML schemas and translators has greatly facilitated this evolution in representations.

We have come to recognize the collaborative nature of content development in the TLCTS mode. The original implementation of an authoring tool for TLCTS was based on a concept of a stand-alone desktop tool used by one author at a time—an unrealistic assumption. The latest iterations have embraced the idea of a web application and emphasized features to manage the collaboration process and cater to the specific needs of the different types of users (linguists, artists, production staff, programmers, and so on.)

Application Use and Payoff

Tactical Iraqi was first deployed in June of 2005. Since then, several expanded and improved versions have been released, and additional courses have been developed. Courses are available in multiple versions for different user communities: for U.S. Marine Corps, U.S. Army, and non-U.S. military forces and civilian aid workers.

The courses are used by individuals training on their own and as part of organized training pro-

Kona	big island	**Honua**	world
Hilo	a Polynesian navigator	**Keaka**	theater
Hua (hua'olelo)	word	**Wele (puna welewele)**	(spider) web
Waihona	library	**Uku (ukulele)**	flea
Lapu	ghost	**Paheona**	art
Huli	search	**Hoahu**	warehouse
Kahu	administrator		

Glossary of Hawaiian Terms.

grams. Anyone with a .mil e-mail account can register and download copies either for their own use or for installation in computer labs. In January 2008, a typical month, there were 910 downloads of Tactical Iraqi, 115 downloads of Tactical Pashto, and 146 downloads of Tactical French.

Patterns of usage depend upon availability of training time and suitable computers. Many learners who download copies for their own use study them to completion; depending upon the course, this can require 200 or more hours of training. For military unit training, 20 to 40 hours of training are more the norm due to conflicting demands on training time. Some units combine TLCTS training with classroom training, whereas others rely exclusively on TLCTS for their language and culture training.

Several independent evaluations of Tactical Iraqi and Tactical Pashto have been performed by the U.S. military, the Canadian Forces, and the Australian Defence Force. Results from several rigorous evaluations have been reported in Surface and Dierdorff (2007) and Johnson and Wu (2008). Surface and Dierdorff studied several subject groups: 268 military advisors training at Fort Riley, Kansas; 113 members of the Seventh Marine Regiment; and 8 trainees at the Defense Language Institute Foreign Language Center (DLIFLC). All groups trained for a total of 40 hours, either with Tactical Iraqi alone or a mixture of Tactical Iraqi and classroom instruction. All showed significant gains in knowledge of Arabic language and culture and greater self-confidence in communicating in Arabic. The greatest learning gains were achieved by the DLIFLC group, which trained exclusively with Tactical Iraqi and followed our recommended program of instruction. Six out of 8 participants achieved an ILR proficiency level of 0+ after 40 hours of training.

The marines in the Seventh Marine Regiment were subjects of a Marine Corps Lessons Learned study. The Third Battalion, Seventh Marines (3/7 Marines) returned from their most recent tour of duty in Iraq in December of 2007. The battalion had assigned two members of each squad to undertake 40 hours of Iraqi Arabic language and culture training prior to deployment. The experience of the 3/7 Marines was particularly noteworthy because the battalion did not suffer a single combat casualty during its tour of duty. To understand the role of Tactical Iraqi in the Seventh Marines' success, members of the two battalions were asked to complete questionnaires and the officers in charge of the 3/7 were interviewed. The questionnaires are still being tabulated, but transcripts of the officer interviews are available, and their comments are strikingly positive. The marines who trained with Tactical Iraqi were able to perform many communicative tasks on their own, without reliance on interpreters. This enhanced the battalion's operational capability, enabled the battalion to operate more efficiently, and resulted in better relations with the local people. This provides indirect evidence that Tactical Iraqi contributed to a Kirkpatrick level 4 training effect (that is, impact on job performance) (Kirkpatrick 1994). Follow-on assessments of the 3/7 Marines' language retention are planned for this year.

Future Work

We continue to improve TLCTS based on experience gained with earlier versions of TLCTS courses. Current work includes broadening the architecture to support reading and writing skills and deepening the curricula and platform to help learners attain higher levels of language proficiency. Language proficiency training is particularly challenging because it helps trainees get to the point where they can construct and understand new sentences in the target language, which complicates speech processing.

We continue to develop our web-based Wele client as an option for learners with less powerful computers. The increased ease of installation and use will hopefully compensate for the reduced lev-

el of three-dimensional rendering and interaction that will result from the constraints of web applications.

We are adapting TLCTS virtual-human technologies so they can be integrated into mission rehearsal training simulations. The training simulations will be populated with nonplayer characters that speak the target language. This poses new challenges for authoring since military trainers with no special technical training will create their own virtual worlds and populate them with nonplayer characters.

We continue to develop new courses and develop pilot versions of future courses. A Chinese course for college and high school Chinese programs is currently being developed in collaboration with Yale University. A pilot Cherokee game, intended to help preserve Native American language and culture, is being developed in collaboration with Thornton Media, Inc.

Acknowledgments

Notes

1. Many of the architecture components have Hawaiian names. Hawaii, a cultural crossroads, is evocative of Alelo's mission to promote cross-cultural interchange. A glossary of Hawaiian terms is included in this article.

2. www.adobe.com/products/flex/

3. developer.apple.com/ipod

4. htk.eng.cam.ac.uk

5. julius.sourceforge.jp/en_index.php

6. www.mindmakers.org/projects/BML

References

Baker, C.; Fillmore, C.; and Lowe, J. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. San Francisco: Morgan Kaufmann Publishers.

Corbett, A., and Anderson, J. R. 1995. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction* 4(4): 253–278.

Feng, D., and Hovy, E. 2006. Learning to Detect Conversation Focus of Threaded Discussions. Proceedings of the Human Language Technology/North American Association of Computational Linguistics Conference (HLT-NAACL 2006). New York: Association for Computing Machinery.

Johnson, W. L., and Wu, S. 2008. Assessing Aptitude for Language Learning with a Serious Game for Learning Foreign Language and Culture. *Proceedings of the Ninth International Conference on Intelligent Tutoring Systems*. Lecture Notes in Computer Science 5091. Berlin: Springer.

Kirkpatrick, D. L. 1994. *Evaluating Training Programs: The Four Levels*. San Francisco: Berrett-Koehler.

Mateas, M., and Stern, A. 2005. Structuring Content in the Façade Interactive Drama Architecture. In *Proceedings of the First Artificial Intelligence and Interactive Digital Entertainment Conference*. Menlo Park, CA: AAAI Press.

Meron, J.; Valente, A.; and Johnson, W. L. 2007. Improving the Authoring of Foreign Language Interactive Lessons in the Tactical Language Training System. Paper presented at the SLaTE Workshop on Speech and Language Technology in Education, Farmington, PA, 1–3 October.

Mote, N.; Johnson, W. L.; Sethy, A.; Silva, J.; and Narayanan, S. 2004. Tactical Language Detection and Modeling of Learner Speech Errors; The Case of Arabic Tactical Language Training for American English Speakers. Paper presented at the InSTIL/ICALL 2004 Symposium on Computer Assisted Learning, Venice, Italy, 17–19 June.

Samtani, P.; Valente, A.; and Johnson, W. L. 2008. Applying the SAIBA Framework to the Tactical Language and Culture Training System. Paper presented at the First Functional Markup Language Workshop, AAMAS 2008, Estoril, Portugal, 13 May.

Si, M.; Marsella, S. C.; and Pynadath, D. 2005. THESPIAN: An Architecture for Interactive Pedagogical Drama. *Proceedings of the Twelfth International Conference on Artificial Intelligence in Education*. Amsterdam, The Netherlands: IOS Press.

Surface, E., and Dierdorff E. 2007. *Special Operations Language Training Software Measurement of Effectiveness Study: Tactical Iraqi Study Final Report*. Tampa, FL: U.S. Army Special Operations Forces Language Office.

Traum, D., and Hinkelman, E. 1992. Conversation Acts in Task-Oriented Spoken Dialogue. *Computational Intelligence* 8(3): 575–599.

Vilhjalmsson, H., and Marsella, S., 2005. Social Performance Framework. In Modular Construction of Humanlike Intelligence: Papers from the 2005 AAAI Workshop. Technical Report WS-05-08, Association for the Advancement of Artificial Intelligence, Menlo Park, CA.

W. Lewis Johnson is cofounder, president, and chief scientist of Alelo Inc. Prior to that he was a research professor in computer science at the University of Southern California/Information Sciences Institute. Alelo realizes his vision to promote the learning of foreign languages and cultural competency worldwide. Alelo's game-based learning environments are in widespread use by military trainees in the United States and other countries. This work has been recognized by multiple awards, including the 2005 DARPATech Significant Technical Achievement Award, the 2007 I/ITSEC Serious Games Challenge, and the 2008 Los Angeles Technology Council Award. Alelo is now partnering with Yale University to develop integrated suites of learning materials for Chinese and other languages and is developing web-based learning materials for distribution worldwide by Voice of America. Johnson holds an A.B. in linguistics from Princeton University and a Ph.D. in computer science from Yale University. He is a member of the steering committees of the International Artificial Intelligence in Education Society, the International Conference on Intelligent User Interfaces, and the International Foundation for Autonomous Agents and Multi-Agent Systems.

Andre Valente is cofounder and CEO of Alelo Inc. Alelo's game-based learning environments are in widespread use by military trainees in the United States and other countries. This work has been recognized by multiple awards, including the 2005 DARPATech Significant Technical Achievement Award, the 2007 I/ITSEC Serious Games Challenge, and the 2008 Los Angeles Technology Council Award. Prior to Alelo, Valente worked as an executive for software startup companies, managed software development, and consulted for businesses in the software, manufacturing, media, and aerospace areas. He also worked as a research scientist at the University of Southern California. Valente received a Ph.D. in computer science from the University of Amsterdam and an MBA from the University of Southern California. He has published three books and more than 50 technical articles on knowledge management, knowledge systems tools. and business process management.