Scaling up in Speaking Proficiency by Supporting Robust Learning Behaviors[1]

Alicia SAGAE[a,b], Rohit KUMAR[a,b] and W. Lewis JOHNSON[a]
[a] *Alelo, Inc., 11965 Venice Blvd., Los Angeles, CA 90066 USA*
[b] *Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA*

Abstract. As language learners scale up in terms of speaking proficiency, they expand the volume of expressions in the target language that they can use in a *robust* fashion, combining vocabulary and syntactic structures into novel constructions and transferring them to new contexts. In the ISLET Project, sponsored by the Office of Naval Research (ONR), we have developed a strategy for allowing a dialog-rich language and culture training system to keep up with this expressive growth. The strategy comprises metrics, methods, and software tools aimed at achieving the necessary increase in coverage to support language learners at an ACTFL Intermediate High speaking level.

Keywords. Language Proficiency, Scalability, Authoring Tools, Language Learning, Mini-Dialogs, Utterance Templates

Introduction

For a language learner, scaling up in terms of proficiency requires increasing the variety of lexical, syntactic, and pragmatic structures that he or she can comfortably use. Consider some of the differences among levels in the ACTFL Proficiency Guidelines for Speaking [1]. While Novice-level speakers are able to "satisfy a very limited number of immediate needs," Intermediate-level speakers have broader range of vocabulary, covering "simple personal needs and social demands to survive in the target language culture." Novice-level speakers are most comfortable with declarative sentence structures that allow them to "respond to simple questions," while Intermediate-level speakers are able to expand their syntactic range to include "obtain[ing] and giv[ing] information by asking and answering questions." At the Advanced level, the Guidelines explicitly describe a learner who can use "a variety of communicative devices," which stands in contrast to Intermediate-level learners who "direct conversation on generally predictable topics."

These comparisons show us that as the learner gains proficiency, his or her generative capacity grows, resulting in a wider variety of expressions that can be comfortably and efficiently called into use during a conversation. In this paper we discuss the consequences of this phenomenon for a dialog-rich language training system that must keep up with the needs of learners whose proficiency is growing. The most direct consequence is that, as the learner "scales up" in terms of proficiency, the system must scale up in terms of the sheer volume of target language expressions that it

[1] This work is supported by the Office of Naval Research under the ISLET Project.

can teach and respond to appropriately. Addressing this type of scalability is one of the issues being pursued by the ONR-funded ISLET Project, with the objective of developing an engaging computer-based language learning environment that takes trainees up to an ACTFL Intermediate High speaking level. In this work we introduce a strategy, comprising metrics, methods, and software tools, that supports this objective by increasing the volume of learner expressions that are gracefully handled by Alelo's Tactical Language and Culture Training System (TLCTS), an ISLET platform.

1. Targeting Robust Language

To solve the problem described above, training system developers must focus their efforts on adding coverage for expressions that learners are likely to use as they spend more time with the system and become more proficient with its content. To discover what these expressions might be, we examine the relationship between proficiency and *robust learning*.

Referring again to the ACTFL Guidelines, we observe that the hallmark of an Intermediate-level speaker, for example, is that he has mastered Novice-level skills to a point at which he can comfortably expand the range of contexts, both topical and syntactic, in which those skills are used. This observation corresponds closely with the definition of robust learning published by the Pittsburgh Science of Learning Center: "[robust learning is] learning that is retained for long durations, transfers to novel situations, or aids future learning" [2]. Hence, a learner who applies rehearsed phrases in novel contexts (task/context *transfer*), or who applies rehearsed linguistic formulas to generate novel patterns (*constructive* use of language) is demonstrating his proficiency with these language skills by using them in a robust fashion. A system that better supports these attempts at robust language will capture more of the expressive space into which learners expand as they attain higher proficiency.

We present a strategy for addressing robust learner language in Alelo's Tactical Language and Culture Training System (TLCTS) that involves three components:

1. A set of definitions for curriculum features that indicate support for robust learner language
2. A methodology for using these definitions as content-analysis metrics, allowing TLCTS curriculum authors to see where improvements could be made to finished content
3. Tools that increase the output of existing workflows in terms of coverage of target language expressions, so that new content can be generated with increased robustness built into the authoring process

This research is ongoing and we will present preliminary results from components 1 and 2, with a more detailed description of user testing of a new tool that contributes to component 3. Each of these efforts contributes to the overall goal of scaling the system up in terms of target learner proficiency.

2. Language and Culture Curricula in TLCTS

Alelo's Tactical Language and Culture Training System (TLCTS) [3] helps people acquire functional skills in foreign languages and cultures using a serious game platform. TLCTS courses include Tactical Iraqi™ and Tactical French™, among many.

TLCTS employs a task-based approach, where the learner acquires the skills needed to accomplish particular communicative tasks [4], then practices these skills in real-time dialog with conversational virtual humans. Heavy emphasis is placed on spoken communication: learners must learn to speak the target language to complete the lessons and play the games.

Although the most current TLCTS curricula teach advanced language skills, they are organized around libraries of *utterances* that are enumerated in advance by curriculum designers ("content authors"). An utterance is a basic unit of TLTCS content; it constitutes a unique word or phrase that appears in the course of instruction. Some example greetings and introductions from Tactical Iraqi 4™ are given below[2]:

as-salaamu 9aleykum (peace be upon you)	*aani ismi jon* (I'm called John)
wa 9aleykum is-salaam (and upon you peace)	*aani jon* (I'm John)
ismi (name)	*ismi jon* ([I'm] called John)

Some utterances are presented to the user directly during the course of instruction, and others are known to the system but only activated when the learner deviates in a predictable way from what he has been taught. These deviations may be considered recognizable robust language. Deviations that have not been enumerated in advance are outside the scope of the current TLCTS curricula; the learner is not supported in his attempts to uses these variations, since the system is not equipped to provide an appropriate response.

This description indicates one path to increasing support for robust language in TLCTS: by establishing metrics for measuring where an existing course does and does not support robust use of language, we can alert the authors who are driving content development and direct their efforts toward increasing this support.

3. Curriculum Features that Support Robust Language

We have started to explore the set of features that indicate support for robust learning within a TLCTS course. We can classify these features as indicators of language transfer and indicators of constructive use of language. Language transfer indicators are features that allow the learner to re-use an utterance outside of the context in which it was learned. Greetings provide a common example. A greeting may be introduced in Lesson 1, in a concrete practice dialog that can be easily memorized. When the same greeting appears in Lesson 10, in a new and more complex dialog setting, this is an opportunity for the learner to demonstrate transfer of the utterance by removing it from the context in which it was rehearsed and applying it in a new context. The appearance of the same utterance in multiple contexts is a feature of the curriculum that allows the learner to make this demonstration.

Other features may indicate that the curriculum supports the constructive use of language skills. For example, a learner who has been introduced to *aani ismi jon* (my name's jon) and *ismi jon* (name's john) may observe the utterance *aani min il-mariinz* (I [am] from the Marines). In subsequent exercises, he may generate *min il-mariinz* (from the Marines), which was not explicitly taught, based on generalization of the

[2] These utterances are displayed using a transparent phonetic transcription that allows Arabic phrases to be rendered in ascii.

pronoun-drop pattern. A curriculum that allows this variation to be recognized supports the learner's attempt at robust language use.

4. Translating Features into Analysis Metrics

We have also performed preliminary experiments to investigate how features of the type described in Section 3 can be operationalized and implemented as analysis metrics for an instance of a TLCTS curriculum. An example is the "N-contextual-variations" feature. This feature estimates robustness using the context in which an utterance is introduced in a TLCTS course. Some utterances are introduced to the learner as explicit objects of instruction. These utterances seem to be core items within the course overall and they occur frequently, once taught. Other utterances appear only as alternative answers in multiple-choice quizzes, or in other contexts where several variants of an expression are presented to the learner together. These utterances seem to be explicit opportunities for learners to practice variations, recombinations, transfer of vocabulary, and other behaviors that indicate robust learning, specifically constructive use of the vocabulary and syntactic patterns involved.

To validate this characterization we created a ranked list of all utterances from Tactical Dari™, where each utterance was assigned a score based on the N-Variations property. Inspection of this list by a proficient student of Dari revealed that utterances with an N-Variations score of 2 or greater superficially correspond to utterances that exhibit constructive use (and therefore robust learning) of language. These would correspond with non-core items, according to our definition above. Further work is needed to establish whether this analysis is valid generally. However the development strategy is consistent with our goal of converting features of the curriculum into implemented metrics; our next step will be to present the results of N-Variations analysis to content authors and allow them to review the lists of "core" and "non-core" utterances. By presenting the content in a new format to authors, we give them the opportunity to add missing non-core variations of core greetings, for example. According to the definitions presented in Section 1, this contributes additional system coverage in the expression space where higher-proficiency learners are likely to grow.

5. New Tools for Authoring Robust Content

The two components introduced so far assist the content author in assessing and then manually improving the support for robust language in an existing TLCTS curriculum. The final component of our strategy is to develop authoring tools that make the existing content-authoring processes more expressive. In addition to letting authors review existing content for robust learning support, we would like to push the improvements upstream in the content-creation pipeline so that new curricula can cover a wider range of learner expressions.

This component of the strategy is complementary to the other two, and we have implemented an example of one such authoring tool that has been deployed for use in the most current TLCTS products. In the remainder of this section we will describe the design, development, and user testing of a new *Mini-Dialog Editor*.

Figure 1. A Mini-Dialog from Tactical French™

5.1. Mini-Dialogs in TLCTS

A mini-dialog comprises a two-turn spoken exchange between the learner and a conversational virtual human. The purpose of a mini-dialog is practice: the learner is exercising the performance of a linguistic behavior he has already learned. However there is pedagogical activity as well, since the exchange is typically followed by feedback from the tutoring agent, a persistent text-based channel for displaying messages to the learner that relate to his performance (e.g. *Good job* or *That was close, but you used the wrong tense*). Figure 1 shows a mini-dialog from Tactical French™. Mini-dialogs allow learners to practice speaking in the target language in a tightly constrained conversational setting. There are about 800 mini-dialogs in Tactical Iraqi™ and over 300 in Tactical French™. They constitute about 15% of the instructional content in TLCTS curricula.

To create a mini-dialog, a content author specifies a textual instruction, e.g. *Tell this person your name*. The instruction may be accompanied by an audio prompt, as shown in **Figure 1**. In this figure, the audio prompt is a recording of the phrase *Qu'est-ce que vous allez enseigner* (What are you going to teach)? The instruction is *Respond by saying "We're going to begin with individual arms."* The learner is expected to give a spoken response.

In order to support feedback to the learner, the content author also enumerates a set of expected responses, in the form of a list of utterances. Each utterance is annotated with a correctness label, indicating whether the utterance is a correct response to the prompt, and a string, representing the feedback that should be given by the tutoring agent if this utterance is used by the learner. Hence, mini-dialog authoring is one example of a curriculum element that depends on the authoring conventions described in Section 2. As we have established, it is critical to expand the coverage of these elements so that learners will continue to receive relevant feedback even as the expressive range of their responses grows. As the learner "scales up" in terms of proficiency, the system must keep up. To achieve this, we introduce a new authoring tool, one that allows authors to generate long lists of alternative expressions without typing each one by hand. Instead, the new mini-dialog editor relies on a compact grammar-based representation called utterance templates.

5.2. Harnessing Utterance Templates

The utterance template (UT) formalism introduced in [5] is a context-free grammar (CFG) syntax similar to HTK syntax. It provides a compact notation for sets of

utterances. The most basic utterance template consists of a string ("coffee") and an optional list of feature-value pairs ("{#drink:coffee} {#hot:true}"). Unlike a formal semantic grammar of the type used in [6], this notation does not include a prescriptive formalism for the semantics of the features or for the granularity of the non-terminals. As a result, the non-terminals in an utterance template can represent linguistic chunks at different levels of granularity ranging from morphemes to words or phrases.

$chunk1 = (va commencer |commencera);
$chunk2 = (par |avec); $chunk3 = (les |des);
$answer = On $chunk1 $chunk2 $chunk3 armes individuelles.; (1)

A more detailed description of the properties of utterance templates and their relationship to other grammar formalisms is given in [5]. In this section we focus on *adoption*: how can these templates be most effectively used by content authors, who are not trained to write context-free grammars, in order to increase the coverage of response sets in mini-dialogs? We propose that the generative power of UTs can be used to create variations from authored responses through the process of *templatization*. For example: By templatizing the response *On va commencer avec les armes individuelles* into the utterance template shown in (1), seven additional variations of the response can be generated which share the same correctness label and feedback.

5.3. The Mini-Dialog Editor

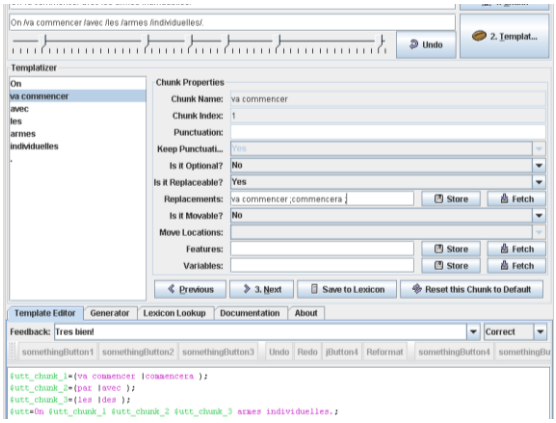

Figure 3a. The Templatizer

To provide access to the expressive power of utterance templates without exposing the mechanics of context-free grammar writing to our content authors, we developed a graphical user interface (GUI) called the Templatizer. The interface is shown in Figure 3a. The templatizer encapsulates meta-rules for CFG grammar writing in the form of *power operations*. These operations translate segments of a response utterance ("chunks") into non-terminals in an utterance template that is being composed by the tool, under the hood. Currently the templatizer has three power operations which were identified by observing common kinds of variations found among the responses of existing mini-dialogs:

- Power Operation 1: Is the chunk *Optional*?
  E.g.: *[Hello.] I am John Smith.* (2)
  Hello is an optional chunk.
- Power Operation 2: Is the chunk *Replaceable*? If so, specify the replacements.
  E.g.: *Hello. (I am |My name is) John Smith.* (3)
  The chunk *I am* is replaceable with *My name is*.
- Power Operation 3: Is the chunk *Movable*? If so, specify the move locations.
  E.g.: (*Thanks.) You are very kind.| You are very kind. (Thanks.)* (4)
  Thanks is a movable chunk that can move to the begin or end locations.

The templatizer is similar to other authoring tools for educational systems [7] that enable both non-programmers and programmers to create complex AI representations. Authoring tools for the Atlas system, for example, also used a grammar-generating GUI to help content authors list possible student responses [8]. However that system required a second pass by a systems engineer to arrive at a grammar. The MD editor, in contrast, is a one-pass tool. As a result the non-terminals of our grammars do not necessarily correspond to domain concepts; a tradeoff that is convenient for our particular application.

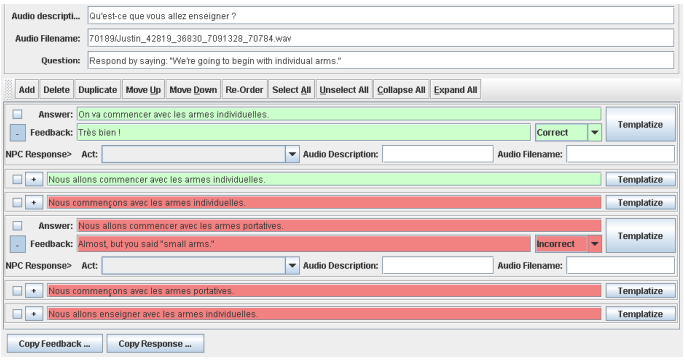

Figure 3b. The New Mini-Dialog Editor

We developed a new authoring interface (Figure 3b) for mini-dialogs that uses the templatizer. To make it easy for authors to transition to the new tool, we imposed a design requirement that the new workflow for creating a mini-dialog should remain as similar as possible to the existing workflow, while still giving authors the opportunity to invoke new features when desired. Authors can still add or edit responses manually.

Clicking on the *templatize* button for a response brings up the templatizer with that response allowing authors to apply power operations on its chunks and create an utterance template. The responses generated by the template can then be exported back to the mini-dialog editor. The exported responses share the correctness label and feedback of the response that was templatized. The intended usage of the templatize button is to allow authors to create variations of responses after they have authored the mini-dialog in a manner similar to the way they author mini-dialogs using the existing editor. Most variations created this way are likely to share the correctness label and feedback. Unintended variations created as a result of over-generation can be deleted.

5.4. Evaluation

We conducted an experiment to evaluate the new mini-dialog editor. One of the main goals of this evaluation was to measure coverage improvements in mini-dialogs that were authored using the new tools. In order to generate results that we could compare to mini-dialogs authored without the new tool, we emulated the existing workflow as nearly as possible. This included the practice of multi-authoring. Before the new tool was introduced, standard practice dictated that a single mini-dialog would be created by one author and passed to two to three additional authors who added to the existing list of responses and corrected errors. In our evaluation we also tracked the interaction of the new tool with multi-authoring.

Four members of Alelo's authoring team who had comparable experience with using the existing mini-dialog editor participated as subjects for this evaluation. The experiment was conducted over three phases. Phase I consisted of a tutorial and practice session. Phase II was spread over three one hour long sessions. During each session the subjects authored a different mini-dialog using the new tools. To compensate for the relatively small number of subjects, we divided each of the sessions of phase II into four sub-sessions referred to as *edits*. During the first edit (20 minutes long) of every session, all four subjects authored the same mini-dialog from scratch. In the second, third and fourth edit (10, 8 and 7 minutes long respectively), the subjects circulate the mini-dialogs authored in the first edit among themselves and made improvements. By the end of each session, all four subjects had a chance to edit the mini-dialogs started by each of them. Phase III consisted of a survey and debriefing.

5.5. Results

We chose precision and recall as outcome metrics. An ideal mini-dialog captures all the possible learner responses, i.e. high recall, and provides accurate feedback for each response, i.e. high precision. To compute precision and recall, we used the set of all responses authored by any author in any of the four edits of each *task* (i.e. one mini-dialog) as an approximation for the set of all possible learner responses for that task. This set consists of both *relevant* and *irrelevant* responses, where relevant responses are useful teaching examples, according to an independent annotation by a subject matter expert in French language instruction. We also report survey responses from Phase III of our evaluation.

	Task1		**Task2**		**Task3**	
	Existing	**New**	**Existing**	**New**	**Existing**	**New**
Relevant Responses	11	70	15	143	5	23
Precision	1.00	0.39	1.00	0.74	0.83	0.27
Recall	0.16	1.00	0.10	1.00	0.22	1.00

Table 1. Task-wise precision, recall, and no. of relevant responses

We observed that the mini-dialogs authored using the new editor had 13 to 17 times more responses in their response sets compared to existing mini-dialogs (authored without the new tool) for the same task. Using the new tool, authors can create a large number of responses in a short amount of time. There were about 5 to 10 times more relevant responses in the new mini-dialogs, as summarized in Table 1. The low precision of new content suggests that the new tools improve the coverage of responses at the cost of introducing some irrelevant ones. Figure 4 shows a plot of

precision and recall metrics for each mini-dialog authored during the evaluation as well as the existing mini-dialogs. An ANOVA on the F-measure using the task, subject and edit as factors revealed a significant effect of task ($F(2,47)=20.7$, $p < 0.001$) and edit ($F(3,47)=1.8$, $p < 0.001$) on the metric. There is no significant difference between subjects. As the mini-dialog goes through multiple edits, its F-metric improves.

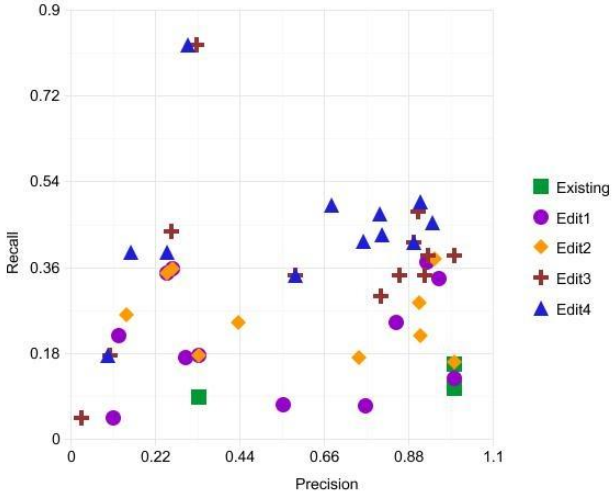

Figure 4. Precision vs. Recall for the various response sets

On the survey, all subjects indicated that the new mini-dialog editor was helpful. Three subjects responded that the quality of the mini-dialogs authored using the new tools were much better and one responded that it was about the same. However, three subjects suggested that they would prefer not to use the new tool when authoring very simple mini-dialogs. There was a mixed response regarding perceived task completion rate while using the templatizer. Two subjects thought the rate was about the same, one thought it was much faster, and another thought it was slower: the templatizer generated so many responses that it became time-consuming to ensure that all of them were correct and appropriate.

6. Conclusions

For a language learner, the process of gaining proficiency is one of scaling up. New vocabulary and linguistic structures are acquired, and new ways of combining them lead to growing expressive power. These combinations are closely related to demonstrations of robust learning of linguistic skills. In order to stay relevant to the learner's needs, a training system has to scale up its coverage of the target language accordingly. If we fail to address them, the system will be unable to respond appropriately when the learner exhibits robust learning. These issues are particularly important in systems like Alelo's TLCTS, which supports learner dialog with conversational virtual humans, and on projects like ISLET, which aim to support learners at ACTFL Proficiency levels higher than Novice level.

Our strategy for scaling up linguistic coverage in TLCTS uses the link between proficiency and robust learning to motivate expanded coverage of constructive language, specifically. Next, we define a practice for analyzing course content based

on features that indicate support for constructive language, giving authors a way to review and then to improve existing curricula. Finally, we present a tool that enables new curricula to cover a wider range of learner expressions at a reduced cost in authoring time. Evaluation of the tool indicated a benefit consistent with findings in the literature related to knowledge-authoring tools, including [7], [9], [10], [11], [12]. To further validate the benefit, we need to author and improve additional courses using this strategy and submit these courses to human evaluations.

6.1. Future Work

This work fits into a larger agenda of research at Alelo addressing many dimensions of scalability, in addition to scaling in terms of proficiency. Other dimensions include scaling to large amounts of course content, multiple target languages, and large numbers of learners. These topics are described in [3], [13], and [14].

In the ISLET project, in particular, we will continue to explore ways of supporting higher-proficiency language learners. Allowing a learner to engage in target-language dialog with conversational virtual humans (CVHs) is a feature that sets TLCTS apart in the field of language instruction systems. Research efforts are currently underway that will help us to reduce the number of dialog breakdowns that occur during these engagements. These efforts will build on the work described in this paper by expanding to all stages of the speech processing pipeline: (1) speech-to-phoneme transcription, (2) utterance recognition, (3) utterance interpretation, and (4) intent planning.

References

[1] ACTFL, ACTFL Proficiency Guidelines – Speaking, http://www.actfl.org/, 1999

[2] Pittsburgh Science of Learning Center, Mission Statement, http://www.learnlab.org/about.php, 2009

[3] W.L. Johnson & A. Valente, Tactical Language and Culture Training Systems: Using Artificial Intelligence to Teach Foreign Languages and Cultures, in *Proceedings of IAAI 2008*, March 2008

[4] C. J. Doughty & M. H. Long, M.H. Optimal psycholinguistic environments for distance foreign language learning, *Language Learning & Technology 7(3)*, pp. 50-80, 2003

[5] J. Meron, A. Valente, W. L. Johnson, Improving the Authoring of Foreign Language Interactive Lessons in the Tactical Language Training System, *Workshop on Speech and Language Technology in Education (SLaTE)*, 2007

[6] Paul Thompson, Mark Stairmand, and William Black. (2004) "Utterance Planning in an Agent-based Dialogue System". In Proceedings of the 3rd International Conference on Natural Language Generation, University of Brighton, Brockenhurst, England, July

[7] V. Aleven, B. M. McLaren, J. Sewall, K. Koedinger, K. The Cognitive Tutor Authoring Tools (CTAT): Preliminary evaluation of efficiency gains. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems (ITS 2006)*, pp. 61-70, 2006

[8] Jordan, P., Rose, C. P., and Vanlehn, K. (2001). Tools for Authoring Tutorial Dialogue Knowledge, Proceedings of AI in Education 2001.

[9] CDAC Bangalore, Visual Java Speech Grammar Development Environment, *http://www.ncb.ernet.in/matrubhasha/visualjsgf.shtml*, 2004

[10] S. Channarukul, S. W. McRoy, S. S. Ali, JYAG & IDEY: A Template-Based Generator and Its Authoring Tool, *Proceedings of the 18th National Conference on Artificial Intelligence*, 2002

[11] Y. Y. Wang, A. Acero, Rapid development of spoken language understanding grammars, *Proceedings of the 9th Eurospeech*, 2005

[12] C. P. Rosé, C. Pai, J. Arguello, Enabling Non-Linguists to Author Advanced Conversational Interfaces Easily, *Proceedings of FLAIRS*, 2005

[13] W.L. Johnson & A. Valente, Collaborative Authoring of Serious Games for Language and Culture, *SimTecT 2008*, March 2008

[14] W.L. Johnson, Serious Use, *JAIED, 2009 (in press)*