Methodological Issues in Evaluating a Spoken CALL Game: Can Crowdsourcing Help Us Perform Controlled Experiments?

Manny Rayner, Nikos Tsourakis

University of Geneva, FTI/TIM, 40 bvd du Pont-d'Arve, CH-1211 Geneva 4, Switzerland

{Emmanuel.Rayner,Nikolaos.Tsourakis}@unige.ch

Abstract

We summarise a series of experiments we have carried out over the last three years on CALL-SLT, a speech-enabled web-based CALL game for learning and improving fluency in domain language, focussing on the methodological aspects. In particular, we argue that our previous evaluations have been systematically flawed due to the lack of a control group. We present a detailed description of our most recent evaluation, where 130 subjects, recruited using crowdsourcing methods, followed a short course in basic French over a period of one week, with 24 subjects completing the course. About a third of the subjects (half of the ones that finished) were assigned to a control group who used a version of the system with speech recognition feedback disabled; subjects in both groups demonstrated significant improvements in language skills over the duration of the experiment, but the improvements were significantly larger for the non-control subjects. We argue in conclusion that this type of experiment opens up interesting new ways to attack the difficult problem of performing controlled experiments with CALL applications.

Index Terms: CALL, speech recognition, evaluation, methodology, crowdsourcing

1. Introduction

Many people have now built CALL tools which use speech recognition and other complex technology. But how can we establish that the advanced functionalities incorporated in these tools do anything useful, in terms of actually helping students develop their language skills? In this paper, we present a case study based on a series of evaluations we have carried out on different versions of CALL-SLT [1], a web-enabled CALL app for learning and improving fluency in domain language.

In the specific instance of CALL-SLT, the central idea is to use speech recognition to provide feedback to the student about their ability to speak in a foreign language. The system prompts the student, indicating what they are supposed to say; the student responds, and the system either accepts or rejects their response. But the student also has a much simpler way to learn. If they are stuck, they have the option of asking for help and hearing a correct response, which they can imitate. Listening and imitating is almost certainly useful, so the question is whether the sophisticated speech-recognition functionality adds anything extra. Without some kind of control group, it seems difficult to make any strong claim to this effect.

In the rest of the paper, we explain how we have developed an evaluation methodology designed to address these issues. We start by describing the CALL-SLT system (§ 2) and critically examining the adequacy of our previous evaluations (§ 3). The novel results of the paper are in § 4 and § 5, where we present our most recent experiment, carried out over the Amazon Mechanical Turk (AMT) on 130 crowdsourced subjects, in which we included a control group who used a version of the system where feedback from speech recognition was disabled. We argue in the final section that this style of evaluation can potentially address the methodological difficulties associated with our earlier efforts; the question is the extent to which the new problems it creates are acceptable.

2. The CALL-SLT System

CALL-SLT [1] is an open-source speech-based translation game designed for learning and improving fluency in domain language. It is based on the "spoken translation game" idea originating with [2]; a related application is TLTCS [3]. The system is accessed via a client running on a web browser. Most processing, in particular speech recognition and linguistic analysis, is carried on the server side, with speech recorded locally and passed to the server in file form [4]. The current version, available at `http://callslt.org`, supports French, English, Japanese, German, Greek and Swedish as L2s and English, French, Japanese, Arabic and Chinese as L1s.

The system is based on two main components: a grammar-based speech recogniser and an interlingua-based machine translation (MT) system, both developed using the Regulus platform [5]. This architecture presents several advantages in the context of the web-based CALL task. The system is not related to a particular language or domain, as in [2]. The Regulus platform offers many tools to support addition of new languages and new coverage (vocabulary, grammar) for existing languages: the recogniser's language model is extracted by specialisation from a general resource grammar in order to get an effective grammar for a specific domain, with the specialisation process driven by a small corpus of sentences. The general grammar can thus easily be extended or specialised for new exercises by changing the corpus, enabling rapid development of new content. The specialised grammar-based language models give good recognition performance on in-coverage sentences even without speaker adaptation; for example, the evaluation exercise described in [6] showed that recognition Word Error Rates for native speakers were very low, typically around 1–2%.

Each turn begins with the system giving the student a prompt, consisting of a surface realisation of an interlingua structure in a predicate-argument notation called Almost Flat Functional Semantics (AFF; [7]). In the plain version of the system, the surface realisation is a text string formulated in a telegraphic version of the L1. The student gives a spoken response; it is in general possible to respond to the prompt in more than one way. Thus, for example, in the version of the system used to teach English to French-speaking students, a simple text prompt might be:

```
DEMANDER DE_MANIERE_POLIE BIÈRE
```

("ASK POLITELY BEER"). representing the underlying interlingua representation

```
[null=[utterance_type,request],
 null=[politeness,polite]
 arg2=[drink,beer]]
```

The responses "I would like a beer", "could I have a beer", "please give me a beer", or "a beer please" would all be regarded as valid.

The system decides whether to accept or reject the response by first performing speech recognition, then translating to an interlingua representation, and finally matching this interlingua representation against the interlingua representation of the original prompt. A "help" button allows the student, at any time, to access a correct response, given in both written and spoken form. The text forms come from the initial corpus of sentences or can be created by the MT system to allow automatic generation of variant syntactic forms. The associated audio files are collected by logging examples where users registered as native speakers got correct matches while using the system. Prompts are grouped together in "lessons" unified by a defined syntactic or semantic theme.

The student thus spends most of their time in a loop where they are given a prompt, optionally listen to a spoken help example, and attempt to respond to the prompt. In the version of CALL-SLT used in the present experiment, the system gave the following minimal feedback to the response. First, it echoed back the student's recorded utterance; second, it placed a border around the text part of the prompt which was either green (accept) or red (reject). We have experimented with more complex feedback, but the difficulty of making it sufficiently reliable means that it is perceived by many students as confusing rather than helpful. At each turn, the student can repeat the current prompt, use arrow controls to move to the next or previous prompt, or switch to a different lesson. The interface used is shown in Figure 1.

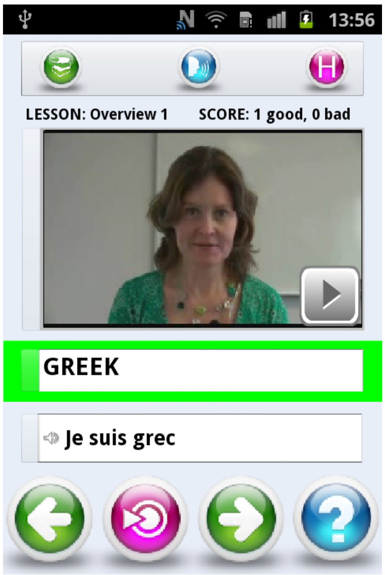

Figure 1: Screenshot of "video" version. The video is saying "Quelle est ta nationalité?" (What is your nationality?); "GREEK" is the text part of the prompt, and "Je suis grec" is a help example.

In the plain version of the system, the relationship between abstract interlingual representations of prompts and their surface text realisations is defined by means of another Regulus grammar [8]; the abstract representation is converted into the prompt by running this grammar in generation mode. In the versions of CALL-SLT used here, this mechanism was extended to generate multimedia prompts. The grammar was modified slightly so that some lexical rules define elements of the form `multimedia:`$\langle$`MultimediaTag`$\rangle$. In a post-processing step, the multimedia tags are removed from the string, and replaced by names of prerecorded multimedia files according to a table which defines a non-deterministic mapping from $\langle$multimedia-tag, lesson$\rangle$ pairs to files. Thus, continuing the previous example, the French multimedia version of the prompt intended to elicit a response similar to "I would like a beer" would be

```
multimedia:ask-for-drink BIÈRE
```

In the initial multimedia configuration we have deployed in the present experiment, multimedia prompts are concretely realised by playing a recorded file corresponding to the multimedia tag and displayed the remaining text; so in the above example, the system would play a recorded video file with a question meaning "What would you like to drink?", while the text "BIÈRE" was displayed.

3. Previous experiments

We start by describing previous evaluations of CALL-SLT which did not use a control group, all of which followed the same basic pattern. A group of students were asked to use the system; we then analysed logged data, looking for evidence that the students improved their language skills between the start and the end of the period in question. Some evaluations were performed over short periods, ranging from a day to a week; others over longer spans of time, as part of a formal language course.

To take a typical example, [9] reported an experiment where ten students used the French-for-Chinese-speakers version for two sessions over one day. We argued that the results provided evidence that subjects had learned from using the system. First, students had a higher proportion of utterances accepted by the system in the later utterances than in the earlier ones, this difference being statistically significant. Second, grammar and vocabulary tests carried out before and after the experiment showed large differences; most of the students appeared to have picked up some vocabulary, and there was also reason to believe that they had consolidated their knowledge of grammar.

Looking critically at the design, we can advance various objections against the validity of our conclusions. One obvious question is whether the fact that students have more utterances accepted by the system after they have used it for a while really does mean that they have improved their generative spoken language skills. Other explanations are a priori quite possible. In particular, they may only have become more skillful at using the interface, learning to speak in a way that is better adapted to the machine, but not necessarily better in itself. The experiments described in [6], however, suggest that these criticisms are not so serious. When native speaker judges are presented with pairs of utterances chosen so that both utterances are responses by the same student to the same prompt, one of which is accepted by the system and one rejected, they tend to agree reasonably well with the recogniser about which member of the pair is better.

A more serious objection, however, is that, even if the results unambiguously show that the student has improved their language skills over a given period, it is still not clear that the improvement can be ascribed to the fact that the student has been using the system. This problem is particularly acute when use of the system is integrated into a formal language course; given that the student is also receiving other kinds of instruction, it is obviously possible that any improvement measured is independent of use of the system. Even if the student is only learning through use of the system, at least over the duration of the experiment, it is still unclear which aspects of the system are responsible for the improvement. In an application like CALL-SLT, the student spends a large part of their time listening and repeating, which may well be helpful for them. It remains to be shown that any of the more sophisticated system functionalities are useful in practice.

Considerations like those above naturally point in the direction of performing controlled experiments, where students using the system are contrasted against a suitable control group. Unfortunately, experience has shown that it is far from easy to define such a group, partly because motivation is always an important factor in language learning. For example, suppose, as in e.g. [10], that we pick subjects randomly from one class, assigning half of them to the group using the system and the other half to the control. The two groups of students will talk to each other. If the system is perceived as useful, which the authors claim in the cited study, it is reasonable to wonder whether students in the control group felt correspondingly unmotivated; it is methodologically better if no subject is aware that any version exists except the one they are using.

If, on the other hand, we take the two groups from two different classes that have no contact with each other, not mixing them, it is impossible to know whether the classes are comparable. Most teachers we have asked say their experience suggests high variability between classes. Yet another possibility is to use a crossover methodology, letting students in the same class alternate between the two groups. Some clear successes have been claimed for this methodology, in particular by the LISTEN project [11, 12, 13]; if the learning effect from using the system is large enough, as appears to be the case there, it is reasonable to hope for a clear result. There are however many known problems with crossover, since it is difficult to account correctly for the effect of using the main system and the control version in different orders. In the context of CALL, students may once again be disappointed if they like the main system and are then forced to use the inferior control, and react accordingly.

For the kinds of reasons outlined above, it has often been argued that controlled experiments are unproductive in CALL [14], and that single-case design methodologies [15] are more appropriate. Recently, however, the introduction of easily available crowdsourcing platforms like the Amazon Mechanical Turk (AMT) has opened up new possibilities. In a large, diverse online community, it is not unreasonable to hope that subjects can be chosen randomly, and in general have no contact with each other; under circumstances like these, a controlled experiment has greater chances of avoiding the known methodological pitfalls. In the next two sections, we describe an experiment of this kind carried out on CALL-SLT.

4. Controlled evaluation using crowdsourcing

The experiment we describe here was carried out in early 2013 using a multimedia-enabled Android phone version of the French CALL-SLT system. The main content consisted of four lessons, *about-me* (simple questions about the subject's age, where they live, etc); *about-my-family* (similar questions about family members); *restaurant* (ordering in a restaurant) and *time-and-day* (times and days of the week). Three additional lessons called *overview-1*, *overview-2* and *revision* will be described shortly. The course was designed for students with little or no previous knowledge of French. It covered about 80 words of vocabulary and a dozen or so basic grammatical patterns.

We created four different versions of the basic system. Three of them differed only in the way the multimedia part of the prompt was realised: in **video** it had the form of a recorded video segment of a human speaker, in **avatar** it was an animated avatar, and in **text** it was a piece of text. The fourth version, **no-rec**, was the same as **video**, except that the student was given no feedback to show whether speech recognition and subsequent processing had accepted or rejected their response.

Subjects were recruited through AMT; we requested only workers from the US. After discovering during a previous study that experiments of this kind can easily attract scammers, we required all workers to have a track record of at least 50 previously completed Human Interface Tasks (HITs), at least 80% of which had been accepted.

The experiment was carried out in two cycles, each of which had the same sequence of eight HITs. In the first HIT, the task was to check that one version of the app (we chose **no-rec**) could be successfully run on an Android phone. Subjects who gave a positive response were then randomly assigned to the four different versions of the system and given different versions of the subsequent HITs. AMT "qualifications" were used so that subjects doing one version of a HIT were unable to see that HITs for other versions existed. The seven HITs were issued at 24-hour intervals; workers were paid $1.00 for the first HIT and $2.00 for each subsequent one, reasonable pay by AMT standards. The HITs had the following content:

Pre-test: The student was asked to do *overview-1* and *overview-2*, each of which consisted of a balanced selection of examples from the other lessons. During *overview-1*, they were encouraged to use the Help function as much as they wished, so the main skill being tested was ability to imitate. In *overview-2*, Help was switched off, so the main skill tested was generative ability in spoken French.

Lessons 1–4: The student was asked to attempt each of the four lessons in turn, one lesson per HIT, with Help turned on. They were told to spend a minimum of 20 minutes practising, and speak to the system at least 25 times.

Revision: The student was warned that the next HIT would be a test (they were not told what it was), and was asked to revise by doing the *revision* lesson, which contained the union of the material from the four main lessons, for at least 20 minutes.

Post-test: The student was asked to do *overview-1* and *overview-2* again. They were told that the intent was to measure how much they had learned during the course, and were asked to do the test straightforwardly without cheating.

The purpose of the pre- and post-tests was to measure the progress the students had made during the main course of the experiment by comparing their results across the two rounds. The mode of comparison will be described shortly.

In the first cycle, we started with 100 subjects. The second column of Table 1 shows the number of students left in play after each round of HITs. At the end of the cycle, there were 17 students who had completed both the pre- and post-tests. A preliminary examination of the results suggested that students performed similarly on the three versions which gave recognition feedback, but worse on **no-rec**; there was not, however, sufficient data to be able to draw any significant conclusions.

Round	Remaining	
	Cycle 1	Cycle 2
Recruit	80	22
Pre-test	36	14
About-me	29	11
My-family	24	10
Restaurant	22	9
Time-and-day	20	8
Revision	18	8
Post-test	17	7

Table 1: Number of students left after each round in the two cycles.

We decided that the most interesting way to continue the experiment was to collect more data for **no-rec**; in the second cycle, we consequently started with 30 subjects, assigning all of them to the **no-rec** group. The third column of Table 1 shows the number left after each round. At the end of the cycle, we had adequate data for 12 subjects in **no-rec** and 12 in the union of the three groups which included recognition feedback, which we will call **rec**. The analysis in the next section thus focusses on exploring the difference between **no-rec** and **rec**.

5. Analysis of results

The main hypothesis we wish to investigate when comparing **no-rec** and **rec** is the obvious one: whether including recognition feedback in the application helps the student. Our basic strategy is equally obvious. For each of the two versions, we compare student performance in the pre- and post-tests. We wish to determine whether this difference is significantly larger in **rec** than in **no-rec**.

rec			**no-rec**		
ID	B-S-W	Signif	ID	B-S-W	Signif
1	1 7-13-8	—	13	6-18-3	—
2	4-14-2	—	14	4-13-1	—
3	9-6-1	$p < 0.05$	15	2-18-2	—
4	9-18-1	$p < 0.05$	16	7-15-6	—
5	8-19-0	$p < 0.02$	17	7-19-2	—
6	10-12-5	—	18	14-9-4	$p < 0.05$
7	6-12-1	—	19	18-7-3	$p < 0.01$
8	8-5-0	$p < 0.02$	20	4-22-1	—
9	6-15-3	—	21	5-15-6	—
10	5-14-9	—	22	10-15-2	$p < 0.05$
11	9-12-2	—	23	5-17-6	—
12	12-11-5	—	24	9-17-2	—

Table 2: Improvement between pre-test and post-test for **rec** and **no-rec** versions, broken down by student. "B-S-W" shows the number of prompts on which the student performed BETTER, SAME and WORSE. "Signif" gives the significance of the difference between BETTER and WORSE according to the McNemar test. Students who described themselves as beginners are underlined.

The pre- and post-tests are the same[1] and contain a total of 28 prompts (13 without help available, 15 with). We compare a given student's performance on each prompt by determining whether the system accepts the student's response or not. As already noted, this correlates reasonably with human judgements [6]. Students can get BETTER (not recognised in pre-, recognised in post-), WORSE (recognised in pre-, not recognised in post-), or stay the SAME (identical outcomes in both tests).

We can compare either across students or across prompts. The simplest way to compare across students is to take each student and count how many examples of BETTER/WORSE/SAME (B/W/S) they get. We can then look at the difference between BETTER and WORSE using the McNemar test to find how significant it is (Table 2); note that $B + S + W$ does not always total to 28, since students sometimes omitted a few items from one or both tests. The comparison turns up four students in the **rec** group who get a significant difference, against three in **no-rec**; in the right direction, but obviously not strong evidence that **rec** is better. Other more complex tests also failed to show a statistically significant difference when we compared across all students, though some were close. It is however worth noting that we do get a significant difference on the two-tail t-test ($t = 1.7$, $df = 11$, $p < 0.01$) when we use only the subset of students, underlined in the table, who described themselves as beginners.

Prompt	BETTER-SAME-WORSE, score			
	rec		**no-rec**	
With help				
P1	**7-1-1**	**66.7**	4-7-1	25.0
P2	**2-8-0**	**20.0**	3-7-3	0.0
P3	1-6-0	14.3	**5-5-2**	**25.0**
P4	**1-7-0**	**12.5**	1-8-4	–23.1
P5	3-3-3	0.0	**4-7-2**	**15.4**
P6	**5-6-3**	**14.3**	5-7-4	6.2
P7	4-3-5	–8.3	2-7-3	–8.3
P8	**3-2-2**	**14.3**	2-5-3	–10.0
P9	**3-2-2**	14.3	**6-7-3**	**18.8**
P10	4-4-1	33.3	**5-5-1**	**36.4**
P11	**4-6-1**	**27.3**	3-5-4	–8.3
P12	**6-6-2**	**28.6**	4-7-2	15.4
P13	**3-4-0**	**42.9**	4-6-2	16.7
Without help				
P14	**3-8-0**	**27.3**	4-7-1	25.0
P15	**3-7-1**	**18.2**	1-10-1	0.0
P16	3-5-2	10.0	**4-6-0**	**40.0**
P17	4-4-3	9.1	3-6-2	9.1
P18	1-10-0	9.1	**2-9-0**	**18.2**
P19	5-4-1	40.0	**6-5-1**	**41.7**
P20	**2-8-0**	**20.0**	3-7-2	8.3
P21	**6-3-2**	**36.4**	3-6-2	9.1
P22	**4-6-2**	**16.7**	2-7-1	10.0
P23	**2-7-0**	**22.2**	1-9-1	0.0
P24	**3-7-1**	**18.2**	1-8-1	0.0
P25	2-7-1	10.0	**2-9-0**	**18.2**
P26	3-7-1	18.2	**2-8-0**	**20.0**
P27	**7-4-0**	**63.6**	7-5-0	58.3
P28	**3-6-0**	**33.3**	3-9-0	25.0

Table 3: Improvement between pre-test and post-test for **rec** and **no-rec** versions, broken down by prompt. The version with the larger improvement is marked in **bold**.

Comparing across prompts produces a convincing result even when we use all the students (Table 3). This time, we look at all the B/W/S scores for a given prompt and version, using the measure $(B - W)/(B + W + S)$. The value will be 100% if every example is BETTER,

[1] We wondered if it was methodologically sound to use the same items for the pre- and post-tests. Students were however going to take the two tests at least a week apart, during which they would practice many similar examples. We felt it was unlikely that they would remember the specific sentences from the pre-test, and that it was more important to give ourselves the option of performing a clear item-by-item comparison.

zero if BETTER and WORSE are equal, and –100% if every example is WORSE.

We can now perform a prompt-by-prompt comparison of **rec** and **no-rec**, contrasting the scores. For example, looking at prompt P11, we have under **rec** $B = 4$, $S = 6$ and $W = 1$, giving a score of $(4-1)/(4+6+1) = 3/11 = 27\%$. Under **no-rec**, we have $B = 3$, $S = 5$ and $W = 4$, giving a score of –8.3%. Applying the Wilcoxon signed-rank test to the whole set of prompts, the comparison between **rec** and **no-rec** on the above measure yields a difference significant at $p < 0.02$.

6. Conclusion and discussion

We can reasonably argue that the experiment described above shows that speech recognition actually does help the student improve their speaking ability. Nonetheless, the modest size of the difference compared to the **no-rec** control group makes us rather thoughtful; in particular, some control students showed significant improvements. Unfortunately, the result leaves it uncertain whether several of our earlier experiments can be interpreted as showing that the student improvements we found need be due to any interesting properties of the system.

We find some aspects of the crowdsourced evaluation methodology attractive, but we are so far reluctant to make strong claims for it. On the negative side, there are several obvious weaknesses: the one which concerns us most is the fact that we have very little control over our subjects. When working with ordinary students, we have the opportunity to meet them, and we usually have some idea of their motivation for wanting to use the CALL tool. (For most of our experiments, we have used subjects who were already learning the language in question). Here, we recruit people randomly through a crowd-sourcing site, and their motivation is unclear. A fair number of the people who completed the course did appear to be interested in learning French: they left positive comments, and, more significantly, many of them logged sessions which were longer than the 20 minutes we required. But not all of them did this.

The nature of the recruitment process readily explains the fact that only a small proportion of the subjects (24 out of 130, or 18%) reached the end of the 8 HIT series. At the beginning, subjects had no clear picture of the level of involvement required in order to complete the course; they only understood this after they had completed the second HIT. It is unsurprising that many of them decided afterwards that they did not want to spend a substantial part of the next week learning to speak better French. The rate of attrition dropped sharply after the third HIT, when subjects knew what to expect. We could have recruited a larger pool, but were limited by financial constraints; the whole experiment cost about $750, a non-trivial amount in our context.

Despite the known problems, our current feeling is that crowdsourced evaluation is well worth further investigation, and opens up interesting new possibilities for carrying out controlled experiments in CALL; as pointed out by Jurčiček and his colleagues [16], whose experiences seem to be fairly similar to ours, the fact that it enables cheap, rapid recruitment of a diverse pool of users is worth a good deal. We expect to see other researchers experimenting with these techniques.

7. References

[1] M. Rayner, P. Bouillon, N. Tsourakis, J. Gerlach, M. Georgescul, Y. Nakao, and C. Baur, "A multilingual CALL game based on speech translation," in *Proceedings of LREC 2010*, Valetta, Malta, 2010.

[2] C. Wang and S. Seneff, "Automatic assessment of student translations for foreign language tutoring," in *Proceedings of NAACL/HLT 2007*, Rochester, NY, 2007.

[3] W. Johnson and A. Valente, "Tactical Language and Culture Training Systems: using AI to teach foreign languages and cultures," *AI Magazine*, vol. 30, no. 2, p. 72, 2009.

[4] M. Fuchs, N. Tsourakis, and M. Rayner, "A scalable architecture for web deployment of spoken dialogue systems," in *Proceedings of LREC 2012*, Istanbul, Turkey, 2012.

[5] M. Rayner, B. Hockey, and P. Bouillon, *Putting Linguistics into Speech Recognition: The Regulus Grammar Compiler*. Chicago: CSLI Press, 2006.

[6] M. Rayner, P. Bouillon, and J. Gerlach, "Evaluating appropriateness of system responses in a spoken CALL game," in *Proceedings of LREC 2012*, Istanbul, Turkey, 2012.

[7] M. Rayner, P. Bouillon, B. Hockey, and Y. Nakao, "Almost flat functional semantics for speech translation," in *Proceedings of COLING-2008*, Manchester, England, 2008.

[8] P. Bouillon, S. Halimi, Y. Nakao, K. Kanzaki, H. Isahara, N. Tsourakis, M. Starlander, B. Hockey, and M. Rayner, "Developing non-European translation pairs in a medium-vocabulary medical speech translation system," in *Proceedings of LREC 2008*, Marrakesh, Morocco, 2008.

[9] P. Bouillon, M. Rayner, N. Tsourakis, and Q. Zhang, "A student-centered evaluation of a web-based spoken translation game," in *Proceedings of the SLaTE Workshop*, Venice, Italy, 2011.

[10] B. Coyne, C. Schudel, M. Bitz, and J. Hirschberg, "Evaluating a text-to-scene generation system as an aid to literacy," in *Speech and Language Technology in Education*, 2011.

[11] R. Poulsen, *Tutoring Bilingual Students With an Automated Reading Tutor That Listens: Results of a Two-Month Pilot Study*. DePaul University, Chicago, IL: Masters Thesis, 2004.

[12] K. Reeder, J. Shapiro, and J. Wakefield, "The effectiveness of speech recognition technology in promoting reading proficiency and attitudes for Canadian immigrant children," in *15th European Conference on Reading*, 2007.

[13] G. Korsah, J. Mostow, M. Dias, T. Sweet, S. Belousov, M. Dias, and H. Gong, "Improving child literacy in Africa: Experiments with an automated reading tutor," *Information Technologies and International Development*, vol. 6, no. 2, pp. 1–19, 2010.

[14] J. Kulik, C.Kulik, and P. Cohen, "Effectiveness of computer-based college teaching : A meta-analysis of findings," *Review of Educational Research*, vol. 50, pp. 177–190, 1980.

[15] C. H. Kennedy, *Single-Case Designs for Educational Research*. Allyn and Bacon, 2005.

[16] F. Jurčiček, S. Keizer, M. Gašić, F. Mairesse, B. Thomson, K. Yu, and S. Young, "Real user evaluation of spoken dialogue systems using Amazon Mechanical Turk," in *Proceedings of Interspeech 2011*, Florence, Italy, 2011.