ISCA Archive
http://www.isca-speech.org/archive

Speech and Language Technology in Education (SLaTE2011)
Venice, Italy
August 24-26, 2011

Evaluation of a Mobile Language Learning System Using Language-Neutral Prompts

Nikos Tsourakis, Manny Rayner, Pierrette Bouillon

ISSCO/TIM/ETI, University of Geneva, Switzerland
{Nikolaos.Tsourakis, Emmanuel.Rayner, Pierrette.Bouillon}@unige.ch

Abstract

We describe two versions of a prototype system for computer assisted language learning hosted on a mobile device. In both versions, the student is given prompts by the machine, which they then have to render in the L2. The two versions differ with respect to the modality of the prompt: one presents it as an L1-based text string, while the other one presents it in pictorial form. The two versions were tested on a group of 32 high school students as part of the Geneva University Student Week. Our findings suggest that male students found the pictorial version significantly easier to use, while female students preferred the text version.

Index Terms: mobile spoken language learning systems, gender differences in HCI, text vs. images in HCI

1. Introduction

As the world becomes smaller, more and more people find it important to learn foreign languages. Although classroom-based course and self-study allow students to acquire basic grammar and vocabulary, conversational interaction with native speakers is needed to acquire real language proficiency. Computer Assisted Language Learning (CALL) systems can bridge the gap between the need and the availability of this type of interaction.

Over the last few years, several academic [1], [2], [3] and commercial [4], [5] CALL systems have appeared, which in various ways use speech recognition. All of these systems use some variant of the obvious strategy: the system gives the student a prompt, suggesting what they are supposed say, and the student then realizes the content of the prompt in the L2. There is, however, a great deal of possible variation, both in the system's strategy for presenting the prompt, and in the user's range of options for realizing it. At one extreme [4], [5], the system tells the student exactly what sentence they are supposed to say, and simply grades them on their pronunciation; at the other [1], the user is presented with a simulated situation, as part of a game, and has to decide on an appropriate utterance. Our system follows [2], and steers an intermediate course. The prompt specifies abstractly what the user is supposed to say, and the student may then realize this as they choose in the L2.

[2] implements this idea in a straightforward way. The prompt is a sentence in the L1; the student is expected to translate it into the L2 and speak it aloud. Speech translation software converts both the L1 prompt and the student's L2 speech into language-neutral representations, which are then compared with each other to give the student feedback on the correctness or otherwise of their response.

A clear drawback of systems like [2] is, however, that presenting the prompts in L1 has the undesirable effect of tying L2 language too closely to L1 in the student's mind. Moreover, studies on vocabulary acquisition suggest that the addition of both visual and auditory features to a text improve comprehension [6]. According to [7], glossing between L1 and L2 using visual and text together is generally better than either alone.

Our system, CALL-SLT [8], builds on the ideas stemming from [2], and in particular reuses speech translation software developed under other projects. We have, however, attempted to address the problems inherent in presenting prompts in the L1. The two versions of the system described here differ with respect to the strategy used. In both versions, prompts are direct renderings of the language-neutral (interlingua) representation. In the first version, these renderings are textual; in the other, they are pictorial.

The version of CALL-SLT we used for the experiments was designed for French students learning English, and hosted on a mobile device. The system uses a restaurant domain, and centers on language for ordering food, reserving a table, asking for the bill, and so on. Vocabulary is around 500 words. The evaluation, which is based on experiments with 32 students during their visit at the University of Geneva, focuses on the question of how textual and pictorial prompts correlates with potential gender HCI issues [9]. Consistent with known tendencies [10], female subjects clearly preferred textual prompts, and male subjects pictorial ones.

The paper is organized as follows. In Section 2 we give an overview of the system, and in Section 3 we present the experimental design. Section 4 presents the evaluation results. The final section concludes.

2. System Description

CALL-SLT's architecture is described in [8] which we briefly summarize, highlighting aspects which relate to the user interface and the presentation of prompts. The system leverages earlier work on Regulus, a platform for building systems based on grammar-based speech understanding [11] and MedSLT, an interlingua-based speech translation framework [12], to develop a generic CALL platform centered on the "spoken translation game" idea.

The game that forms the basis of CALL-SLT is as follows. The system is loaded with a set of possible prompts, created by translating the development corpus into the language-neutral interlingua. Each turn starts with the student asking for the next prompt. The system responds by showing them a textual or pictorial representation of the underlying interlingua for the sentence they are supposed to produce in the L2.

The student decides what she is going to say, presses the "recognize" button, and speaks. The system performs speech recognition, translates the result into the interlingua, matches it against the underlying interlingua representation of the prompt, gives the student feedback on the match, and adjusts the level of difficulty up or down considering the length of the prompt. If the match was successful, the student's recorded speech is also saved for future use.

Figure 1: *Text and Image versions of the system.*

The versions of CALL-SLT used here were deployed on a mobile device, using the architecture described in [13] and the gender independent commercial Nuance Toolkit. Figure 1 shows screenshots for the two versions, illustrating the different types of prompt used. In the "Text" version on the left, the prompt is a French-based string; in the "Image" version on the right, it is a sequence of graphical icons. Each icon is of size 145x145 pixels so the screen of the 800x480 pixels can accommodate 12 of them along with the GUI's buttons. The student may ask for help at any time, which can be of two possible kinds. If the student clicks on the "help" button (bottom right in Figure 1), they are played a correct L2 audio file for the current example, recorded by an L2 native speaker; the system also shows the example in written form. In the image version, the student can also tap on each image and get an English word or phrase corresponding to it.

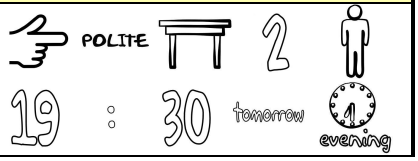
Table 1 summarizes the distribution of 336 images according to their content, which is based on whether or not it contains pure sketch, pure text or both (hybrid). For our domain it is difficult to model everything with pure sketches.

Table 1. *Image type distribution.*

Sketch	Hybrid	Text	Days - Numbers	Symbols
83	102	67	78	6

In particular, specific food names or places were expressed with hybrid images (e.g. pizza Napolitana, Caesar salad) or pure text (e.g. corner, outside). Numbers, day names and symbols like question mark, colon etc were modeled as text. Table 2 gives examples of text and image prompts.

Table 2. *Examples of prompts.*

L1	*Order politely a pizza napolitana*
TEXT	COMMANDER DE_MANIERE_POLIE PIZZA_NAPOLITANA
IMAGE	
L1	*Ask politely to pay in euros*
TEXT	DEMANDER DE_MANIERE_POLIE PAYER EUROS
IMAGE	
L1	*Reserve politely a table for two people at seven thirty tomorrow evening*
TEXT	RÉSERVER DE_MANIERE_POLIE TABLE 2 PERSONNE(S) 19 H 30 DEMAIN SOIR
IMAGE	

3. Experimental Setup

Our evaluation was carried out on 32 students, who tried out the system during the University of Geneva Student Week. The experiments were conducted over a 10-day period, with participants from different schools in the region of Geneva. Subjects were between 16 and 19 years old and had intermediate to advanced English proficiency. All students were native French speakers none of whom were bilingual in English.

The group of participants was split equally between 16 male and 16 female subjects. Each student filled out a demographic questionnaire and was given a 10 minute general presentation of the system, after which they used it for 15 minutes or 30 turns, whichever was shorter. Upon completion they were asked to fill an evaluation questionnaire. Sessions took place in an office environment with moderate background noise (~70 dB), as there were other students interacting with the desktop version of our system at the same time.

4. Evaluation

We divide the presentation of the results into two sections, concerned with objective and subjective results respectively. The evaluation is mostly based on comparative measurements contrasting male and female participants for both text and image versions.

4.1. Objective evaluation

Efficiency. In the objective evaluation we seek to quantify interaction patterns for the users. We counted the average number of speech recognition turns per minute and the average percentage of successful interactions, presented in Table 3. Performance is roughly uniform across configurations and test groups. For example, the mean number of speech recognition interactions per minute for male subjects was 3.2 in the text configuration and 3.3 for the image one. In correlation with the percentage of successful interactions, depicted also in Table 3, we can conclude that neither of the two versions imposes additional effort to the two user groups.

Table 3. *Interaction statistics.*

	♂ Text	♂ Image	♀ Text	♀ Image
turns/min	3.2	3.3	2.9	3.1
% good rec	60.4	57.4	55.9	56.2

In Figure 2 we present plots of cumulative game score against number of turns, for two specific cases where we encountered differences in the configurations and the subject groups. In plot 1 we observe that female subjects initially perform better with the text configuration, although they eventually catch up with the image configuration after 23 interactions (marked by the circle). We observed a different score pattern for male subjects, where they performed equally well in both configurations throughout the session (due to space limitations this plot was not included in Figure 2).

Another suggestive result is presented in plot 2, where we initially see narrow but clear dominance of the male score pattern over the female one in the image configuration, with convergence after approximately 22 interactions (marked by the circle). For the text configuration, both groups perform equally well (again not depicted in Figure 2). Although the mean values in the plots are not statistically significant, they provide a direction for the analysis that follows.

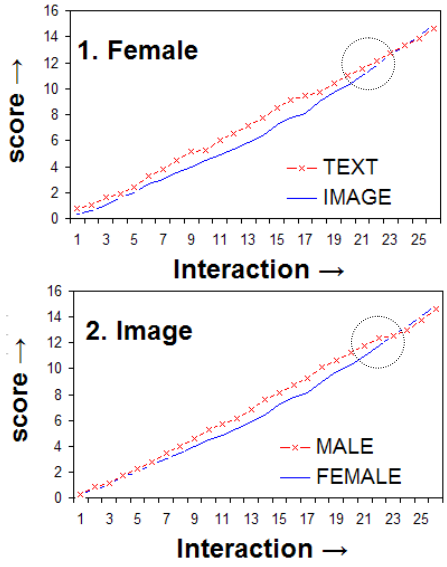


Figure 2: *Score patterns.*

Mental Workload. We quantify mental workload in terms of "thinking time", which we define as the time between the presentation of the prompt (text or images) and the press of the recognition button. During this interval the user can think what to say, press the help button one or more times, or press on one of the available images. The interval includes the announcement time of the help prompts. We define the "first interaction time" as the time spent by the user processing the input and requesting help for the first time or pressing the recognition button. The decomposition of the time intervals is shown in Figure 3.

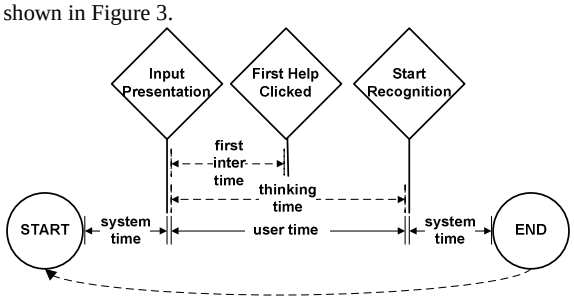


Figure 3: *Decomposition of user/system time.*

We note that, as a user can repeat a specific example more than once, (e.g. if there is a misrecognition in the first trial), the "thinking time" doesn't include the proceeding trials. The first plot in Figure 4 represents the average "first interaction time" for both versions. We can deduce that users tend to interact more rapidly when introduced to the image input. This is probably due to the fact that they know the grammatical structure of the input, which most often resembles the pattern: "I would like" + <object>, and thus focus on the corresponding image of the object. In the "text" configuration they are forced to read the whole sentence before interacting. Although the processing of the image input is probably done sequentially, the human mental model analyzes previous image sketches faster and focuses on the new ones. We used one tail t-test assuming equal variance to calculate statistical significance. On average it takes 3.1 sec longer (4.0 sec to 7.1 sec) ($t=1.91$, $df=14$, $p<0.04$) to process text than image input for male subjects, whereas 1.3 sec ($t=1.68$, $df=14$, $p<0.045$) for the female group.

The average "thinking time" is depicted in the second plot of Figure 4, which was uniform in all configurations and equal to around 9.5 sec. This is consistent with the results shown earlier (Table 3), where male and female subjects have almost the same number of interactions per time for all configurations.

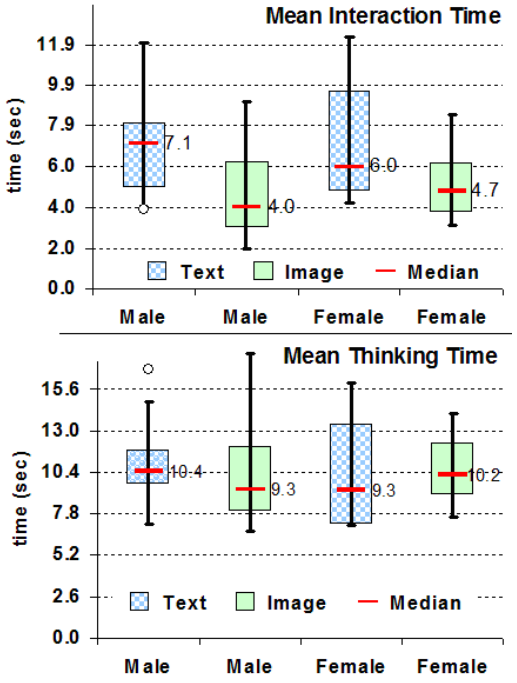


Figure 4: *"Mean time" box-plots (the colored boxes contain the 50% of the samples closest to the mean).*

In Figure 5, we present the average number of help and image clicks per interaction. For example, for the male group we observed an average of 0.6 clicks in the "text" configuration and 1 click (help and image) in the image one.

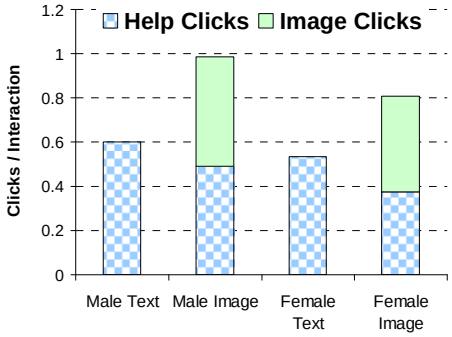


Figure 5: *Help/Image clicks per interaction.*

The frequency of help clicks is not reduced much between the configurations, which suggest users feel more confident listening to a prompt presented by a native speaker. The number with the image configuration is greater for the male group compared to the female one, which may explain to some extent their better than average score performance. By clicking the help button users were forced to listen to the whole help prompt even though they may only need help for a specific

word. On the other hand listening to a native speaker can enhance pronunciation skills. The fact that the clicking of the help button takes place almost the same amount of time in all configurations reveals that users are equally exposed to the correct pronunciation. Image help has also an impact on the out-of-vocabulary range (OOV) presented in Table 4, which is reduced for both male and female subjects in this version.

Table 4. *OOV rates*.

	♂ Text	♂ Image	♀ Text	♀ Image
OOV (%)	3.75	1.79	5.31	2.59
	(t=2.13, df=14, p<0.03)		(t=1.96, df=14, p<0.03)	

Waveform Analysis. We examined another aspect of the problem using the software program praat [14] to calculate the mean intensity and the mean of the second formant frequency of each utterance. For the female subjects we get a mean intensity of 51.06 db in the text configuration whereas 48.62 db in the image one (t=22.46, df=87514, p<0.001). It can be deduced that female users feel less confident in the second case and thus speak less loud. We believe that the correlation between confidence and loudness is positive when someone uses the system in front of others and especially when the evaluator is present. For comparison basis the mean intensity of the native speaker who recorded the help prompts is 56.97 db. The corresponding values for male students are 49.45 db (text), 48.53 db (image) (t= 8.45, df=79777, p<0.001) and 57.43 db (native). In Figure 6 we present the cumulative distribution function (CDF) of the mean intensity for the female subjects, approximated using kernel density functions.

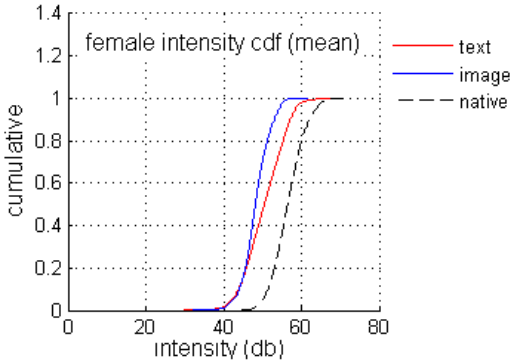


Figure 6: *CDF of the waveforms intensity.*

The mean of second formant related to accent [15], seems to alter across configurations. For females users we get 1620.55 Hz for the text configuration and 1633.47 Hz for the image one (t=9.92, df=112330, p<0.001), whereas the native speaker has a second formant equal to 1624.88 Hz. The values for male students are 1582.63 Hz (text), 1579.14 Hz (image) (t=2.45, df=107999, p<0.01) and 1578.55 Hz (native).

4.2. Subjective evaluation

Upon completion of the experiment each subject had to fill an exit questionnaire. We received reactions ranging between neutral to positive concerning the overall impression towards the system. Based on their answers we present the most interesting results on a 5-point Likert scale, with 5 being the highest (Table 5). For the statement "The game was too difficult" male subjects provided higher score for the text version (2.5) compared to the image one (1.62). For female subjects we observe the opposite pattern. Concerning the issue of how it easy was to pose a question, male users seem more confident in the image version (4.37) rather than the text version (3.25). Again, the female group presents the opposite reaction. The difference in perception of each group towards the system has already been recorded from the objective measurements in previous subsections.

Table 5. *Subjective evaluation score.*

The game was too difficult			
	TEXT	IMAGE	Significance test
♂	2.5	1.62	(t=2.7, df=14, p<0.02)
♀	1.62	2.25	(t=2.54, df=14, p<0.02)
I understood easily what I had to say in English			
	TEXT	IMAGE	Significance test
♂	3.25	4.37	(t=2.02, df=14, p<0.04)
♀	4.25	3.25	(t=2.25, df=14, p<0.03)

5. Conclusions

In this work, we have addressed usability issues in two versions of a prototype system for language learning on a mobile platform. The objective and subjective measurements show that there is a small but clear gender difference regarding the relative usability of the two configurations. Male subjects appear, on average, to have a preference towards images, which they process more quickly than female subjects do; female subjects have the reverse profile.

Other future plans for the prototype include addition of further dialogue processing to support structured language courses and richer multimedia prompts (e.g. audio, video, animated, etc). As stated in [16], gender differences in end-user computing still receive little research attention; therefore any new features of the system should also be evaluated in this respect.

6. References

[1] Johnson, W.L., "Serious use of a serious game for language learning". Artificial Intelligence in Education. 2007.

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

[3] Bernstein, J., et al, "Subarashii: Encounters in Japanese Spoken Language Education". CALICO Journal 16 (3), 361-384, 1999.

[4] Rosetta Stone, http://www.rosettastone.com/

[5] Tell Me More, http://www.tellmemore.com/

[6] Chun, D. and Plass, J., "Effects of multimedia annotations on vocabulary acquisition". Modern Language Journal, 1996.

[7] Chun, D., "CALL technologies for L2 reading". Calling on CALL, 2006.

[8] Rayner M. et al, "A Multilingual CALL Game Based on Speech Translation". LREC, 2010.

[9] Beckwith, L. and Burnett, M., "Gender: An important factor in end-user programming environments?" IEEE. Symposium on Visual Languages and Human-Centric Computing, 2004.

[10] Caplan, P. J., et al., "Gender differences in human cognition", Oxford Univ. Press, New York, 1997.

[11] Rayner, M., Hockey, B.A. and Bouillon, P., "Putting Linguistics into Speech Recognition". CSLI Press, 2006.

[12] Bouillon, P. et al, "Many-to-Many Multilingual Medical Speech Translation on a PDA". AMTA, 2008.

[13] Tsourakis, N., Georgescul, M., Bouillon, P., and Ranyner, M., "Building Mobile Spoken Dialogue Applications Using Regulus". LREC, 2008.

[14] Boersma, P. and Weenink, D., Praat: Doing phonetics by computer. http://www.praat.org/.

[15] Yan, Q. et al., "Analysis, Modeling and Synthesis of Formants of British, American and Australian Accents". ICASSP, 2003.

[16] Beckwith, L., Burnett, M. and Shraddha, S., "Gender and End-User Computing". Human Computer Interaction, Information Science Reference, Hersey – New York, Vol. I, 2009.