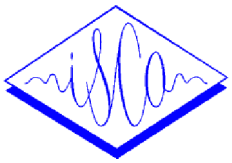
ISCA Archive
http://www.isca-speech.org/archive



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

A Student-Centred Evaluation of a Web-Based Spoken Translation Game

Pierrette Bouillon[1], *Manny Rayner*[1], *Nikos Tsourakis*[1], *Qinglu Zhang*[2]

[1] University of Geneva, ETI/TIM/ISSCO,
40 bvd du Pont-d'Arve, CH-1211 Geneva 4, Switzerland
[2] Faculty of European Languages and Cultures, Guangdong University of Foreign Studies,
Guangzhou, 510420, P. R. China

{Pierrette.Bouillon,Emmanuel.Rayner,Nikolaos.Tsourakis}@unige.ch,
aluzhang@163.com

Abstract

We present an evaluation of CALL-SLT, a web-deployed speech-enabled platform for improving fluency in a limited domain, based on the "translation game" idea of Wang and Seneff. The evaluation used 10 Chinese-speaking students of French, who spent an average of about three hours each practising on a set of five lessons covering elementary and intermediate grammar topics in a restaurant domain. We found significant improvements in student performance, measured both in terms of their ability to be recognised correctly by the application and according to written vocabulary and grammar tests.

Index Terms: CALL, speech recognition, evaluation, Web, French, Chinese

1. Introduction

In this paper, we present a concrete evaluation of CALL-SLT [1], a platform for language learning based on a spoken translation game, which is intended to help a second language (L2) learner to improve fluency in a domain (restaurant, hotel booking, etc). The idea behind a translation game was originally suggested by Wang and Seneff [2]: the learner receives as input a set of L1 sentences (prompts) that have to be verbalised in the L2 language. These sentences are extracted from a list of example sentences defined by the teacher.

In CALL-SLT, we innovate in two ways compared with Wang and Seneff's work. First, the system does not show the learner an L1 sentence, but rather an L1 gloss of the meaning of the sentence. For example, if the L2 is French and the L1 is English, a gloss for "Je voudrais la soupe" would be `ORDER POLITELY SOUP`, a gloss for "Auriez-vous une bouteille d'eau" would be `ORDER POLITELY BOTTLE WATER`, etc. This avoids the undesirable effect of linking too closely the L2 language to the L1 in the student's mind. The focus is thus more on L2 language production rather than translation. A second innovation is that CALL-SLT includes a powerful mechanism to build lesson plans. This mechanism makes it possible to structure automatically the initial set of sentences into fine-grained lessons that pick out subsets of sentences based on predefined lexical, syntactic or semantic properties [3]. The teacher can in this way build exercises that involve specific speech acts (ordering something, asking for something, etc), semantic fields (food, drink, etc) or syntactic structures (questions, conditional tense, etc).

We evaluate a set of French lessons that have been designed, using the lesson plan mechanism, to teach fluency in a restaurant language domain. The experiment was carried out using Chinese-speaking students of French. Students spent about three hours working with the system, and explored five lessons. The focus was on investigating how much they improved at the level of pronunciation, vocabulary and syntax. In the sequel, we begin by briefly presenting CALL-SLT (Section 2). The main body of the paper then describes the experiment (Section 3) and the results (Section 4).

2. The CALL-SLT System

CALL-SLT is an open-source speech-based translation game designed for learning and improving fluency in domain language. The system is accessed via a normal web browser using a Flash interface that can be downloaded in a few seconds; all heavy processing, in particular speech recognition and language processing, is carried on the server side, with speech recorded locally and passed to the server in file form. The current version focuses on the restaurant domain; there are multiple versions, supporting French, English, Japanese and German as L2 and English, French, Japanese, German, Arabic and Chinese as L1.

The system is based on two components: a grammar-based speech recogniser and an interlingua-based machine translation (MT) system, both developed using the Regulus platform [4]. In order to check whether the sentence pronounced by the learner is correct or not, the system first performs speech recognition. The MT system then determines if the recognised sentence corresponds to the meaning of the prompt presented to the learner. To do this, it transforms the sentence into the meaning (interlingua) representation and matches it against the representation of the prompt. The level of difficulty is adjusted up or down depending on whether matching was successful. A help button allows the student, at any time, to access a correct sentence in both written and spoken form. The text forms come from the initial corpus of sentences or can be created by the MT system (this allows automatic generation of variant syntactic forms, which is particularly important for French question constructions). The associated audio files are collected by logging examples where users registered as native speakers get correct matches while using the system.

This architecture presents several advantages for a CALL application. 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 is extracted by specialisation from a general resource grammar in order to get an effective grammar for a specific domain. The specialisation process is driven by a small corpus of sentences,

Figure 1: CALL-SLT client running in a web-browser window, showing the L1 gloss, help, recognition result and lesson help file. The system is freely available for online use at `http://callslt.org`.

constructed so as to contain at least one example of each required word and grammatical structure. This general grammar can thus easily be extended or specialised for new exercises by changing the corpus.

The grammar-based recognition approach is well suited to the web-based CALL task; in particular, it gives good coverage on in-coverage sentences even without speaker adaptation or training data. It is also very rare for recognition to produce ungrammatical sentences, which could give misleading feedback to students. Finally, the interlingua-based MT allows us to produce a language-independent meaning for the sentence which can easily be glossed for different L1s; the Chinese-L1 version used here was adapted from the French one in just a couple of days. In summary, the approach appears to be appropriate to a limited-domain multilingual system that can be accessed by a wide variety of casual internet users.

Figure 1 illustrates the web-based interface used for the experiment. It offers five main functionalities: 1) choosing a language pair and specific lesson (upper left buttons); 2) getting a prompt in the L1 language (middle right button); 3) responding to spoken input (upper right button); 4) getting written and spoken help for a given prompt (lower right button) and 5) getting lesson help (far right panel). Each lesson help file contains written material associated with the lesson, in particular explaining the intended way of speaking and the grammar topic. This is described further in the next session.

3. Experimental setup

The experiment, an extended version of an earlier exercise [5], used 10 Chinese-speaking computer science students who were spending an exchange year in Tours, France. The students had previously done between one and two years of French in China. They had all spent five months in France when we performed the experiment.

We asked the students to use the French-for-Chinese version of the system, loaded with the five lessons shown in Table 1. Each lesson is organised around a particular speech act and grammar topic. The first three lessons are tightly focussed on specific grammatical patterns, while the fourth and fifth are more complex. In particular, the fourth lesson introduces both new vocabulary and also French question constructions, which are generally experienced by beginner/intermediate students as challenging. The fifth lesson recapitulates all the constructions used in the preceding lessons. Each lesson has an accompanying lesson help file, written in the L1 language (here Chinese), which explains the content. A typical lesson help file contains about a third of a screen of text.

The students took part in two sessions. During the first, which was of approximately two hours' duration, they were encouraged to use the system freely in order to practise the content of the lessons. The help function was switched on, and all the students accessed it often, typically on half to two-thirds of the prompts. During the second session, about an hour long, the students went through the same lessons, but this time with the help function disabled. The students were told in advance that the second session would be treated as a test of what they had learned in terms of speaking skills, grammar and vocabulary.

The students were also given a short written test at the beginning of the first session, and then took it a second time at the end of the experiment. The test was divided into sections for vocabulary and grammar/structure. In the vocabulary part, they were given 29 Chinese nominal phrases whose French counterparts occurred in the lesson material, and asked to translate them

ID	#Sents	Topics	Typical examples
1	14	Ordering, conditional tense, singular nouns	Je voudrais la soupe
2	9	Ordering, future tense, plural nouns	Je prendrai des asperges
3	5	Ordering, "s'il vous plaît"	L'agneau, s'il vous plaît
4	46	Requesting, yes-no questions	Auriez-vous un couteau?
5	16	Booking a table, numbers	Je voudrais une table pour trois personnes

Table 1: Lessons used in experiments: lesson ID, number of examples, topics and a typical example.

into French. The grammar/structure part consisted of five sentences to translate and eight "fill-in-the-gap" questions, which between them covered all the grammar topics in the lessons. At the end of the final session, the students were also asked to complete a short questionnaire.

4. Results

We now present the results of analysing the session logs and the written tests. We extract two types of statistics from the session logs: the proportion of successful recognition events, and the proportion of successfully attempted prompts. We consider a recognition event to be successful if the system scores it as a match (as described in Section 2, this does not in general require a word-for-word match of the recognition result against a reference sentence, but is carried out at the semantic level). A prompt is deemed to have been successfully attempted if the student eventually gets a successful recognition event for it, possibly after more than one try. Thus the proportion of successful recognition events can reasonably be considered a measure of the student's ability to pronounce well. When help is switched off, the proportion of successful prompts is a rough measure of what the student has remembered in terms of vocabulary and grammar.

The main results are summarised in Tables 2 to 6; we group the results both by lesson and by subject. In order to be able to make clear comparisons of like with like, we split the first session (with help switched on) into two equal halves. This lets us contrast recognition performance across the two halves, to get some impression of how much students' pronunciation has improved by the halfway mark.

Lesson	Help			NoHelp
	1st half	2nd half	All	
1	26.1	33.6	29.8	39.5
2	24.0	37.7	30.5	48.3
3	12.2	16.5	14.5	29.7
4	23.9	22.6	23.4	22.0
5	22.6	19.3	18.2	23.4

Table 2: Percentage of successful recognition events, grouped by lesson. For the first three columns, subjects had access to online spoken help.

Looking first at the per-lesson recognition table (Table 2), we find substantial improvements for Lessons 1, 2 and 3, both between the first and second halves of the session with help, and even from the session with help to the session without. All the differences are significant at $P = 0.015$ or better according to the Fisher test (two-tailed), except the one for Lesson 3 between the first and second halves of the help session. These exercises are simple and repetitive, and the most likely explanation is that the students are improving their pronunciation and fluency. It is worth noting, in contrast, that there is no improvement in Lesson 4, the one that contains the most complex grammar. When we group the data by student (Table 3), we found that three subjects, 2, 4 and 9, improved their recognition scores significantly ($P = 0.01$ or better) from the help session to the no-help session.

ID	Help			NoHelp
	1st half	2nd half	All	
1	35.8	45.7	41.6	52.8
2	17.3	31.9	24.6	55.8
3	36.3	27.2	31.6	31.9
4	7.3	9.1	8.2	15.6
5	47.4	62.8	54.1	58.8
6	26.7	31.1	28.7	18.4
7	10.6	13.7	12.1	12.9
8	24.4	18.6	21.0	14.6
9	16.1	22.2	19.2	32.7
10	23.0	27.7	25.4	26.9

Table 3: Percentage of successful recognition events, grouped by subject. For the first three columns, subjects had access to online spoken help.

Lesson	Help			NoHelp
	1st half	2nd half	All	
1	87.7	89.2	88.7	88.3
2	82.3	87.3	85.1	75.0
3	59.1	63.6	60.4	72.5
4	82.4	80.2	81.4	65.1
5	67.0	60.4	63.7	71.3

Table 4: Percentage of successfully answered prompts, grouped by lesson. A prompt was considered as successfully answered if the student eventually gave a response which resulted in a successful recognition event. For the first three columns, subjects had access to online spoken help.

Moving now to the statistics for successful prompts, Table 4 presents the results grouped by lesson. The most interesting contrast is between the with-help and without-help sessions; for all of the lessons except the fourth one, there was no significant decrease in average performance at $P = 0.05$. In fact, for lessons 3 and 5 we measured a marginal increase. This accords with the earlier tables, which suggest that most students had mastered the vocabulary and grammar during the first session. The fourth lesson, as already noted, was perceived as clearly harder than the others, and there was indeed a very significant drop in prompt score here ($P = 0.001$).

Grouping the data by subject (Table 5), we find that four subjects (1, 3, 6 and 8) performed significantly worse in the

no-help session, while two (2 and 5) did significantly better; differences for the remaining subjects were not significant. This again agrees well with the recognition data; none of the three subjects (2, 4 and 9) who continued to improve their recognition performance in the no-help session did significantly worse in terms of successful prompts.

ID	Help			NoHelp
	1st half	2nd half	All	
1	80.4	84.8	**81.9**	**53.7**
2	80.0	88.3	**82.9**	**91.9**
3	87.0	87.0	**85.8**	**31.9**
4	55.2	46.3	50.7	42.6
5	95.9	83.7	**90.0**	**97.4**
6	80.6	80.6	**81.5**	**62.8**
7	63.0	66.7	64.8	63.4
8	55.4	51.8	**53.5**	**36.0**
9	84.3	87.1	85.5	94.3
10	85.0	84.0	84.7	78.3

Table 5: Percentage of successfully answered prompts, grouped by subject. A prompt was considered as successfully answered if the student eventually gave a response which resulted in a successful recognition event. For the first three columns, subjects had access to online spoken help. Differences between the with-help and no-help sessions significant at $P = 0.05$ or better are marked in **bold**.

ID	Vocabulary		Syntax		Average	
	Before	After	Before	After	Before	After
1	31	76	69	92	50	84
2	14	79	38	77	26	78
3	24	76	77	100	51	88
4	0	72	15	54	8	63
5	34	83	100	85	67	84
6	3	14	69	46	36	30
7	3	52	15	85	9	69
8	31	72	54	85	43	79
9	21	76	46	92	34	84
10	17	76	61	69	39	69

Table 6: Results of written vocabulary and syntax tests (percentages) administered before and after the sessions with CALL-SLT.

Finally, Table 6 presents the results of the written tests. All but one of the students demonstrated solid improvements on both parts.

5. Summary and conclusions

Table 7 summarises how much subjects improved during the course of the experiment. The written test shows that, at a minimum, all but one of them improved their vocabulary and grammar. In terms of speaking skills, the top three students (2, 4 and 9) achieved significant improvements; since this continued into the session where help was not available, we can reasonably suppose that it includes both spoken vocabulary and pronunciation. The fourth- to sixth-ranked subjects probably improved as well, though the difference was not statistically significant at $P = 0.05$. There is a noticeable correlation between the results of the written and spoken evaluations, and all three of the students who achieved significant improvements in recognition performance also made gains of over 50% in the written test.

ID	Recognition			Written
	1/2	2/NoHelp	1/NoHelp	
2	+14.6	+23.9	+38.5	+52
1	+9.9	+7.1	+17.0	+34
9	+6.1	+10.5	+16.6	+51
5	+15.4	–4.0	+11.4	+17
4	+1.8	+6.5	+8.3	+56
7	+3.1	–0.8	+2.3	+60
10	4.7	–0.8	+3.9	+30
3	–9.1	+4.7	–4.4	+38
8	–5.8	–4.0	–9.8	+36
6	+4.4	–12.7	–8.3	–6

Table 7: Improvements in performance (percentages), by subject: improvement in recognition score from first half of session with help to second half; from second half of session with help to session with no help; from first half of session with help to session with no help; in written tests.

In brief, a plausible guess is that at least half the class learned something, and the top students learned a good deal. This accords with the students' own subjective evaluations in the exit questionnaire. All ten of them considered that interacting with the system had been helpful in terms of improving their French, though only four considered that it allowed them to identify what concrete errors they were making. We feel that this is a fair reflection of the system's strategy, which is based on encouraging students to repeat examples until the recogniser accepts them. Most people who are not already experts appear to improve pronunciation and fluency, though the improvement is largely at an unconscious level. Whether one thinks this is a good thing or not depends largely on one's approach to language pedagogy. The most common negative comment from the students was that recognition was too unforgiving. Since then, we have added a feature to the system which makes it possible to change the level of "recognition strictness". This work will be reported in detail elsewhere.

6. 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, http://www.issco.unige.ch/pub/lrec2010_callslt.pdf.

[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] M. Rayner, P. Bouillon, N. Tsourakis, J. Gerlach, C. Baur, M. Georgescul, and N. Y., "A multilingual platform for building speech-enabled language courses," in *Proceedings of the L2WS Workshop*, Tokyo, Japan, 2010.

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

[5] P. Bouillon, I. Halimi, M. Rayner, and N. Tsourakis, "Evaluating A Web-Based Spoken Language Translation Game For Learning Domain Language," 2011.