Evaluating A Web-Based Spoken Language Translation Game For Learning Domain Language

Bouillon, Pierrette; Halimi Mallem, Ismahene Sonia; Rayner, Emmanuel; Tsourakis, Nikolaos

EVALUATING A WEB-BASED SPOKEN TRANSLATION GAME FOR LEARNING DOMAIN LANGUAGE

Pierrette Bouillon[1], Sonia Halimi[1], Manny Rayner[1], Nikos Tsourakis[1]

[1]Université de Genève, Ecole de Traduction et d'Interprétation (ETI)
Pierrette.Bouillon@unige.ch, Sonia.Halimi@unige.ch, Manny.Rayner@unige.ch, Nikos.Tsourakis@unige.ch

Abstract

We present an evaluation of CALL-SLT, a web-based CALL application based on the "translation game" idea, that can be used for practicing fluency in a limited domain. The version tested was configured to teach basic restaurant French to students whose native language is Arabic. Students spent about an hour and a half each working with the system, and explored between five and eight lessons. The focus was on investigating how they interacted with the system and what they learned most effectively.

Keywords: CALL, speech recognition, machine translation, translation games, French, Arabic

1 INTRODUCTION

In this paper, we present CALL-SLT, a platform for language learning based on a spoken translation game, 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 [4]: the learner receives a set of L1 sentences (prompts) that have to be verbalized 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 to 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 ORDER POLITELY SOUP, ORDER POLITELY BOTTLE WATER, etc. This should avoid undesirable effects 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 [2]. The 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 [1]. The teacher can in this way build exercises that involve specific speech acts (ordering something, asking for something …), semantic fields (food, drink …) or syntactic structures (questions, conditional tense).

In this paper, we evaluate a set of French lessons that have been elaborated, using the lesson plan mechanism, to teach fluency in a restaurant language domain, and describe a concrete experiment carried out using Arabic-speaking students of French at the University of Al Ain, UAE. Students spent about an hour and a half each working with the system, and explored between five and eight lessons. The focus was on investigating how they interacted with the system and what they learned most effectively.

In the sequel, we begin by briefly presenting CALL-SLT (Section 2). The main body of the paper then describes the set of French lessons (Section 3), the experiment (Section 4) and the results (Sections 5 and 6).

2 CALL-SLT, A WEB-BASED SPOKEN TRANSLATION GAME

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 recognizer and an interlingua-based machine translation (MT) system, both developed using the Regulus platform [3]. 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 recognized 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. Depending on whether matching was successful or not, the score is adjusted up or down. A help button allows the student, at any time, to access a correct sentence in both written and spoken form. These sentences come from the initial corpus of sentences or can be generated automatically by the MT system.

This architecture presents several advantages for a CALL application. The system is not related to a particular language or domain, as in [4]. The REGULUS platform offers many tools to support addition of new languages and new coverage (vocabulary, grammar) for existing languages: the recognizer is extracted by specialisation from a general resource grammar in order to get the most effective grammar for a specific domain. The specialisation process is driven by a small corpus of sentences, constructed so as to contain at least one example of each required word and grammatical structure. This general grammar can 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. 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.

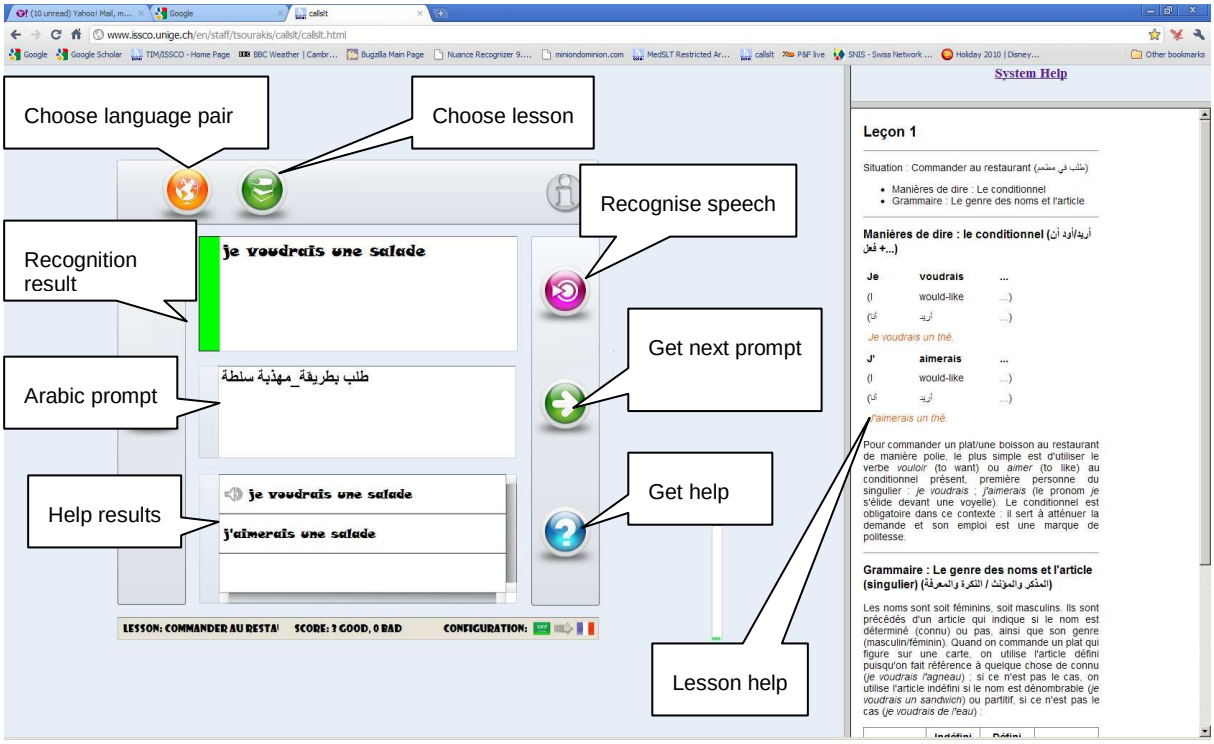

Figure 1. CALL-SLT web interface (version for French L2 and Arabic L1). The student presses the "Get next prompt" button and is shown the Arabic prompt. They then press and hold the "Recognise speech" button and speak, after which the recognition result is displayed. At any time, they can press the "Get help" button and get help examples in written and spoken form; the top two buttons allow them to move to a different language-pair or lesson. The system is freely available for use at http://www.issco.unige.ch/en/staff/tsourakis/callslt/callslt.html. Casual users should log in as "guest" with no password.

Figure 1 illustrates the web-based interface used for the experiment. It offers five main functionalities: 1) choosing a language pair and specific lesson; 2) getting an exercise in the L1 language; 3) responding to spoken input; 4) getting written and spoken help for a given exercise (prompt) and 5) getting lesson help. 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 THE FRENCH LESSONS

The French L2 version of the system contains 23 lessons built using the lesson plan mechanism, and intended to teach fluency in the restaurant domain. Each of them focuses on a specific speech act (*order a dish*, *book a table*, *ask for something*, *pay*, *ask where something is* …), a "way of speaking" (using *conditional*, *future*, *yes-no question, infinitive,* …) and a grammatical topic (*numbers, official time*, …), as illustrated in Table 1.

Id	Speech act	Way of speaking	Grammar topic	Examples
1	Ordering	Conditional	Singular nouns: gender and article	Je voudrais le poulet J'aimerais le poulet
2	Ordering	Future	Plural nouns: gender and article	Je prendrai les fruits
3	Ordering	Conditional (lesson 1) and Future (lesson 2)	Numbers (2-10)	Je voudrais deux cafés J'aimerais deux cafés Je prendrai deux cafés
4	Ordering	Using "S'il vous plaît (SVP)"		Un café SVP Je voudrais un café SVP J'aimerais un café SVP Je prendrai un café SVP
5	Ordering	Conditional (lesson 2) Future (lesson 3) Using "s'il vous plaît (SVP)" (lesson 4)	Complex nouns	Un café au lait SVP Je voudrais un steak de boeuf (SVP) J'aimerais un steak de boeuf (SVP) Je prendrai un steak de boeuf (SVP)
8.	Asking for something	Yes-no questions	Yes-no questions	Auriez-vous un couteau? Est-ce que vous auriez un couteau? Vous auriez un couteau?
9.	Asking for something	Infinitive after a conditional : *j'aimerais avoir …, je voudrais avoir…*	Infinitive	Je voudrais avoir un couteau (*s'il vous plaît*) J'aimerais avoir un couteau (*s'il vous plaît*)

12.	Booking a table for a specific time	Infinitive after a conditional (lesson 9): *j'aimerais réserver une table pour .., je voudrais réserver une table pour…* Yes-no questions (lesson 8)*: Auriez-vous une table de libre pour ...*	Official time	Je voudrais réserver une table pour dix-neuf heures J'aimerais réserver une table pour dix-neuf heures Auriez-vous une table de libre pour dix-neuf heures ?

Table 1: Lessons used in the experiment

The aim of the set of French lessons is to teach both vocabulary and grammar since the learner is forced to try many different structures for the same speech act. For example, lesson 1 focuses on ordering using singular nouns and the conditional tense (*je voudrais …*); lesson 2 also covers ordering, but this time suggest using plural nouns and the future tense (*je prendrai …*), etc. The lessons differ in level of complexity. Some introduce both new vocabulary, ways of speaking and grammar topics (lesson 1), while others focus on one of these points only. For example lesson 6 introduces a new way of speaking (questions) and new vocabulary (thing), while lesson 7 introduces a new way of speaking but uses the same vocabulary as lesson 6. Lesson 5 just recapitulates ways of speaking seen in previous lessons. Each lesson is associated with a help file, as illustrated on the right side of Figure 1, that explains the ways of speaking the student is supposed to use and the current grammar topic.

The aim of the experiment was to evaluate the subset of eight lessons we have just presented.

4 THE EXPERIMENT

The experiment described here was designed to evaluate the set of French lessons described in the preceding section, and was carried out on seven Arabic-speaking students currently studying French at the department of Translation, University of Al Ain, UAE. The subjects were aged 20 to 25, and were at an elementary level in their studies; they had been learning French for one semester, with an average attendance of 4 hours a week. The training was organized during two sessions according to the students' availability. Each session began with a ten-minute presentation during which the students were shown how to use the system. Following this, the student was given a 90 minute slot to practice with it. They were asked to imagine themselves in a restaurant where they would use French to reserve a table, order food and drinks, and ask for other objects like plates, knives, forks or menus, completing as much as possible of the eight lessons outlined in Section 3 above.

We had two main goals. First, we wanted to obtain a rough picture of how the students interacted with the application; in particular, we were interested in finding out whether there were large qualitative differences between the way they approached the lessons, which varied considerably in level of difficulty. Second, we tried to estimate how much they had learned from their 90 minutes of study. We tracked their interaction over the course of the session, and also compared their scores on a simple knowledge test administered before and after they had used CALL-SLT. This test was in two parts, covering vocabulary and grammar respectively. In the vocabulary part, the student was given a list of 31 Arabic words commonly used in the restaurant environment, and asked to find French equivalents. Typical examples are *bottle* (زجاجة), *water* (ماء) and *fish* (سمك). In the grammar exercise, they were asked to construct four sentences, in a polite form, asking for something from the waiter or booking a table. We expected them to use formal structures like *je voudrais .., j'aimerais …, je prendrai …, un/e … s'il vous plaît,* etc*.*. At the end of the session, the subjects completed a brief questionnaire (evaluation grid) to assess their satisfaction and give feedback on the usefulness of the system.

The following sections present the results.

5 HOW STUDENTS USED THE SYSTEM

We start in Table 2 by summarising differences in interaction across the lessons. The four columns are as follows. The first shows the lesson ID, using the identifiers from Table 1. The second shows the average proportion of successful recognition attempts for the lesson, averaged over all the students; thus, for example, `81/247 = 32.8%` in the first line means that there were a total of 247 recognition attempts for lesson 1, of which 81 matched the prompt. The third column shows average success per prompt, where a prompt is considered to have been successfully answered if the student eventually succeeds in responding with a successful recognition, possibly after more than one attempt. The fourth column shows the average number of attempts per prompt for the lesson, and the final one the average number of help events per prompt.

There are several points to notice. First, students almost always used help; except in lessons 3 and 12, they averaged about one help event per prompt. In the case of lesson 3, the lower number of help events is probably explained by the fact that the lesson concentrates on numbers from 2 to 10, which were already inside most subjects' vocabulary, and recapitulates vocabulary and structures already seen in previous lessons.

Looking at the average prompt scores, we get the impression that the first five lessons were definitely appropriate to the students' level of competence. They were on average able to respond correctly to more than half of the prompts, typically within a couple of attempts. It is less clear that lessons 8 and 12 were appropriate; students were averaging only around 10% correct recognitions for these exercises, and succeeded on around a quarter of the prompts. Some students failed to get any examples correct on these lessons, though the best ones achieved credible scores. There were several possible reasons: these lessons introduced challenging new grammatical structure (subject/verb inversion) and new vocabulary, and also occurred at the end, when the students were getting tired and having trouble concentrating. Lesson 9 appeared to be intermediate in level of difficulty, perhaps in part because it reused vocabulary introduced in lesson 8.

A general conclusion we draw from this is that it is important to give students sufficiently many examples that they have time to practice new material, enabling them to consolidate the knowledge they have acquired.

Lesson ID	Av. rec. score	Av. prompt score	Attempts/prompt	Help/prompt
1	81/247 = 32.8%	78/125 = 62.4%	1.98	1.18
2	29/62 = 46.8%	29/46 = 63.0%	1.35	1.00
3	44/141 = 31.2%	40/79 = 50.6%	1.78	0.76
4	21/91 = 23.1%	20/36 = 55.6%	2.76	0.94
5	32/96 = 33.3%	32/54 = 59.3%	1.78	1.06
8	17/143 = 11.9%	17/62 = 27.4%	2.31	1.02
9	5/28 = 17.9%	5/14 = 35.7%	2.00	1.14
12	7/66 = 10.6%	7/30 = 23.3%	2.20	0.77

Table 2. Student performance per lesson: proportion of successful individual recognition attempts, proportion of prompts successfully attempted, average number of attempts per prompt and average number of help requests per prompt.

6 WHAT THE STUDENTS LEARNED

We now estimate what the students learned at the level of pronunciation, vocabulary and syntax.

6.1.1 ***Pronunciation improvement***

As we saw in the preceding section, students appeared comfortable with the first five lessons, but found the last three too difficult. Looking just at the five lessons of appropriate level of difficulty, we counted the number of successful recognition events in the first and last halves of each lesson. The results are shown in Table 3

Lesson ID	Av. rec. score	Av. (first half)	Av. (second half)
1	81/247 = 32.8%	44/122 = 36.1%	37/122 = 30.3%
2	29/62 = 46.8%	12/29 = 41.4%	15/29 = 51.7%
3	44/141 = 31.2%	21/69 = 30.4%	22/69 = 31.9%
4	21/91 = 23.1%	8/44 = 18.2%	11/44 = 25.0%
5	32/96 = 33.3%	13/47 = 27.7%	19/47 = 40.4%

Table 3. Proportion of successful individual recognition attempts per lesson, contrasting performance in the first and second halves of the lesson.

In four of the five lessons, the students got a better score in the second half. The quantity of data is too small for this to be statistically significant, but the result is suggestive; students appear to be improving their pronunciation skills a little during the time they interact with the system. In the final section, we outline a large evaluation exercise we are planning to carry out soon, where we will be able to determine if this effect stands up to closer examination.

6.1.2 Knowledge test

As described in Section 4, the students were also given a short knowledge test before and after using the system. Table 4 shows improvement on the vocabulary part. Before the test, students averaged only 2.9 words out of 31, with two students failing to get any word; after, they averaged 12.6 words correct, with 6 as the worst score. If nothing else, students clearly retained a fair amount of new vocabulary.

User Id	Vocab score (out of 31)	
	Before	After
1	11	19
2	1	11
3	2	10
4	0	6
5	3	17
6	3	16
7	0	9
Average	2.9	12.6

Table 4: Score on vocabulary part of knowledge test before and after the session with CALL-SLT. The maximum possible score was 31.

User Id	Grammar score (out of 4)	
	Before	After
1	2	4
2	0	2
3	No data	No data
4	No data	No data
5	0	2
6	0	2
7	0	2

Table 5: Score on grammar part of knowledge test before and after the session with CALL-SLT. The maximum possible score was 4.

Table 5 shows the results for the grammar part. All of the students who completed the test learned at least two constructions, in particular the polite requesting construction using the conditional (*je voudrais…*) which was the most frequently practiced item. In contrast, only one of the students appeared to have mastered the difficult question and infinitive constructions, and the others did not remember them. This agrees with the impression given by the recognition results in Table 2: the last three lessons seem to be too difficult for this group. It also shows a correlation between the recognition results and what is learned.

6.1.3 Qualitative evaluation

At the end of the session, 5 of the 7 students completed a brief questionnaire. The results are summarised in Table 6. Encouragingly, all the subjects felt that the system was teaching them something useful and met their needs.

Questions	**1**	**2**	**3**	**4**	**5**	**6**	**7**
It is useful						**1**	**4**
It meets my needs						**2**	**3**
It does everything I would expect it to do					**1**	**2**	**2**
It is easy to use							**5**
It is user friendly						**1**	**4**
I can use it without written instructions					**1**	**1**	**3**
It is easy to learn to use it							**5**
I would recommend it to a friend							**5**
It is fun to use							**5**
I feel I need to have it							**5**

Table 6: Questionnaire results. 1 = strongly disagree, 7 = strongly agree.

7 CONCLUSIONS AND FURTHER DIRECTIONS

Our first non-trivial evaluation exercise with CALL-SLT was positive. The students enjoyed using the system and were able to learn something even from a short session. One point that stood out was that some lessons were much more effective than others. The first five lessons worked well; the last three, much less so. It appears particularly important to have enough examples to practice on when the topic covered by the lesson is on the challenging side. The accompanying lesson texts should probably also have had more detailed explanations of the grammar in the L1 language, and not in the L2 as was the case for this experiment.

We will bear these points in mind during our next evaluation, which is tentatively scheduled for February 2011. This time, we plan to use about 20 Chinese-speaking students of French, and ask students to use the system during three separate sessions over the space of a week. This should give us a much clearer picture of the extent to which CALL-SLT can be used as a practical teaching aid.

8 REFERENCES

[1] Rayner M., Bouillon P., Tsourakis N., Gerlach J., Baur C., Georgescul M. and Nakao Y, A Multilingual Platform for Building Speech-Enabled Language Courses. Proceedings of the L2WS Workshop, Tokyo, Japan, 2010.

[2] Rayner M., Bouillon P., Tsourakis N., Gerlach J., Georgescul M., Nakao Y., Baur C. A Multilingual CALL Game Based on Speech Translation. Proceedings of LREC, Valetta, Malta, 2010.

[3] Rayner M., Hockey B.A. and Bouillon P. *Putting Linguistics into Speech Recognition*, CSLI, Stanford, 2006.

[4] Wang C. and Seneff S. Automatic assessment of student translations for foreign languages tutoring. Proceedings of NAACL/HLT 2007, Rochester, 2007.