Evaluating Appropriateness Of System Responses In A Spoken CALL Game

Manny Rayner, Pierrette Bouillon, Johanna Gerlach

University of Geneva, FTI/TIM,
40 bvd du Pont-d'Arve, CH-1211 Geneva 4, Switzerland
{Emmanuel.Rayner,Pierrette.Bouillon,Johanna.Gerlach}@unige.ch

Abstract

We describe an experiment carried out using a French version of CALL-SLT, a web-enabled CALL game in which students at each turn are prompted to give a semi-free spoken response which the system then either accepts or rejects. The central question we investigate is whether the response is appropriate; we do this by extracting pairs of utterances where both members of the pair are responses by the same student to the same prompt, and where one response is accepted and one rejected. When the two spoken responses are presented in random order, native speakers show a reasonable degree of agreement in judging that the accepted utterance is better than the rejected one. We discuss the significance of the results and also present a small study supporting the claim that native speakers are nearly always recognised by the system, while non-native speakers are rejected a significant proportion of the time.

Keywords: CALL, speech recognition, web, evaluation, French

1. Introduction and background

People studying a foreign language need to practise four main skills: reading, writing, listening and speaking. It is relatively easy to build mechanical systems that help with the first three, but the fourth is challenging. The increased emphasis on spoken language in education means that the issues involved have been brought more sharply into focus. In Europe, for example, the influential "Common European Framework of Reference for Language" (CEFR; http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf) has led to substantial changes in language teaching methods. Human teachers cannot easily cope with the increased demand for time spent helping students develop productive speaking skills, and the case for developing mechanical aids has become correspondingly stronger.

There are many applications designed to help improve pronunciation: an impressive and well-documented example is the EduSpeak® system (Franco et al., 2010), and some commercial offerings, like RosettaStone and TellMeMore, have become very popular. These systems, however, generally limit themselves to teaching the student how to imitate: the student listens to a recorded sound file, imitates it to the best of their ability, and is given informative feedback. This does indeed help with pronunciation, but it is less clear that it helps improve spontaneous speaking skills.

A more ambitious approach is to design an application where the student can respond flexibly to the system's prompts. The system we will describe in this paper, CALL-SLT (Rayner et al., 2010), is based on the "spoken translation game" idea originating with (Wang and Seneff, 2007); a related application is TLTCS (Johnson and Valente, 2009). The system prompts the user in some version of the L1, indicating in an abstract or indirect fashion what they are supposed to say; the student speaks in the L2, and the system provides a response based on speech recognition and language processing.

The unspoken assumption behind all applications of this kind is that the application responds in an appropriate way, accepting well-spoken utterances more readily than badly-spoken ones. If this is true, then it is plausible that the system will guide the students towards better speaking habits. Conversely, if it is false, then the system will be useless or even harmful, encouraging the student to speak in an unnatural way in order to get recognised.

The goal of the paper is to critically examine the above assumption. We first present some suggestive results, showing that coarse-grained statistics (WER and SER) for a small set of speakers are in rough agreement with intuitive assessments of speaking ability. We then describe a more careful study, where we take logged data from a formal CALL-SLT evaluation exercise and extract pairs of utterances where the same student has responded to the same prompt, choosing the pairs so that one element is accepted by the system and the other is not. We asked native-speaker judges to listen to both files and say which of the two recorded files they consider better. Reassuringly, the result turns out to be positive: although judgements are rarely unanimous, the judges agree with the recogniser much more often than they disagree with it.

The rest of the paper is organised as follows. Sections 2. and 3. give further background on CALL-SLT and the data collection exercise. Section 4. describes how we performed the judging task, Section 5. presents the results and Section 6. concludes.

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 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. The current version, available at http://callslt.org, supports French, English, Japanese, German, Greek and Swedish as L2s and English, French, Japanese, German, 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 Regu-

Figure 1: The version of CALL-SLT (French for Chinese-speakers) used in the main study, showing the L1 gloss, help, recognition result and lesson help file. The system is freely available online at `http://callslt.org`.

lus platform (Rayner et al., 2006). Each turn begins with the system giving the student a prompt, formulated in a telegraphic version of the L1, to which 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 prompt might be:

`DEMANDER DE_MANIERE_POLIE BIÈRE`

(“`ASK POLITELY 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 potentially valid.

The system decides whether to accept or reject the response by first performing speech recognition, then translating to language-neutral (interlingua) representation, and finally matching against the language-neutral representation of the prompt. 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 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. A response which is correct but which does not match the theme of the lesson produces a warning.

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. If the system accepts, they move on to a new prompt; if it rejects, they will typically listen to the help example and repeat, trying to imitate it more exactly. If they are still unable to get an accept after several repetitions, they usually give up and move to the next example anyway. On reaching the end of the lesson, the student either exits or selects a new lesson from a menu.

The 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 (Wang and Seneff, 2007). 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. It is also very rare for recognition to produce ungrammatical sentences, which could give misleading feedback to students; for example, in the small evaluation exercise summarised in Table 1 below, there were no ungrammatical recognition results for any of the 557 transcribed files.

3. The data collection exercise

The data collection exercise, described in detail in (Bouillon et al., 2011), used 10 Chinese-speaking computer science students who were spending an exchange year in Tours, France. The students, who had previously done between one and two years of French in China and spent

Subject	Level	#Utts	WER	SER
Chinese	Beginner	250	49.9	84.4
Rayner	Intermediate	90	7.0	20.0
Gerlach	Native	138	2.2	7.3
Bouillon	Native	79	0.2	1.3

Table 1: Gross speech recognition measures (Word Error Rate and Sentence Error Rate, given as percentages) for a few subjects with differing levels of expertise in spoken French. The first line show the average performance for the Chinese students involved in the main experiment. The remaining ones show results for corresponding samples recorded by the authors of the paper.

five months in France, were asked to use the French-for-Chinese version of the system, loaded with five sample lessons. Figure 1 illustrates the web-based interface used. The students took part in two sessions, totalling about three hours in duration and yielding a total of 5245 recorded spoken interactions. Each spoken response was stored in recorded form, together with meta-data including the associated system prompt.

The Chinese students clearly found the exercise challenging for a number of reasons. French is phonetically and prosodically a difficult language for Chinese speakers, the students had not been studying long, and at first they had trouble using the headsets and the push-and-hold recognition interface. The experiment was moreover carried out in a small room, with all the students sitting close to each other and talking simultaneously. Sound quality on many of the files was extremely poor, with common problems including cutoffs at the beginning or the end of the recording, low volume, and high levels of background noise.

As a result of all these factors, Word Error Rate (WER) was very high; based on a random sample of 250 wav-files which we extracted and transcribed, we estimate it at around 50%. To reassure sceptics, Table 1 presents the results of a short informal study carried out by the authors of the paper, where we compare the students' performance with our own: each of us recorded and transcribed about a hundred examples, covering the same five lessons as those used in the main experiment. The results, though obviously anecdotal, do at least provide reasonable support for two claims. First, the recogniser is capable of delivering excellent performance with speakers who use it correctly; second, despite the fact that all three authors had practised a good deal during system development, there are clear differences in scores. The native French speakers (Bouillon and Gerlach) get an almost perfect recognition result, with an average WER of a little over 1% between them. In contrast, the intermediate-level speaker (Rayner) has a WER of 7%. The difference in performance between the two native speakers may be due to the fact that Bouillon has a Belgian accent and Gerlach a Swiss one; our impression is that the data used to train the acoustic models underrepresents Swiss speakers.

4. Evaluating responses

Despite the many problems referred to above, we were pleased to find during post-experiment debriefing that nearly all the subjects expressed a positive opinion, and thought that interacting with the system had been a rewarding experience which had taught them useful things about spoken French. It was evidently possible, however, that this positive reaction could have been due either to excessive politeness on the students' part or to some kind of placebo effect. We consequently sought objective evidence that the system was giving them useful feedback about the quality of their spoken language. Ideally, we would like it always to accept their speech when it is above a certain threshold, and otherwise always to reject it. This goal is unattainable, but we wished to estimate how closely we approached it.

To this end, we collated the data so as to find cases where a) the same student had responded more than once to the same prompt, and b) at least one example had been accepted, and at least one rejected. For each such group, we randomly selected one recorded file which had been accepted and one which had been rejected, giving us 413 pairs.

As expected, an initial sampling of the data quickly revealed that, in many cases, the most important characteristic was that one or both files had been badly recorded due to the various issues mentioned in Section 3.. We consequently divided judging into two rounds. During the first round, two system experts listened to all the pairs, and marked ones which exhibited recording problems: this accounted for 243 pairs, about 56% of the data.

The remaining 170 pairs were then judged by three French native speakers, all of whom had worked as French language teachers. None of them had previously been associated with the project or knew the exact point of the evaluation exercise: in particular, we were careful not to tell them that each pair consisted of one successful and one unsuccessful recognition match. Judges were asked to mark each pair to say which element, if either, was better in terms of speech (pronunciation/prosody), vocabulary and grammar. Judging was performed using the Amazon Mechanical Turk, with the judges paid a zero fee. This allowed us to distribute work efficiently over the Web and also simplified the task of writing the judging interface, which could be specified straightforwardly in HTML. A screen-shot of the judging interface is shown in Figure 2.

5. Results

The main results are shown in Table 2. For each judge, we list the number of pairs on which they explicitly agree with the system (i.e. the judge considered that the accepted element of the pair was better) and the number where they explicitly disagree (the judge preferred the rejected element). If the judge did not express a preference with respect to any of the specified criteria, we counted the pair as being neither an agreement nor a disagreement. We also list results for aggregated judgements that are "unanimous" (all three judges), "majority" (at least two out of three judges) and "at-least-1" (at least one judge). The first group of lines shows results for all judgements; the second and third consider, respectively, only judgements based on speech quality and only judgements based on language quality.

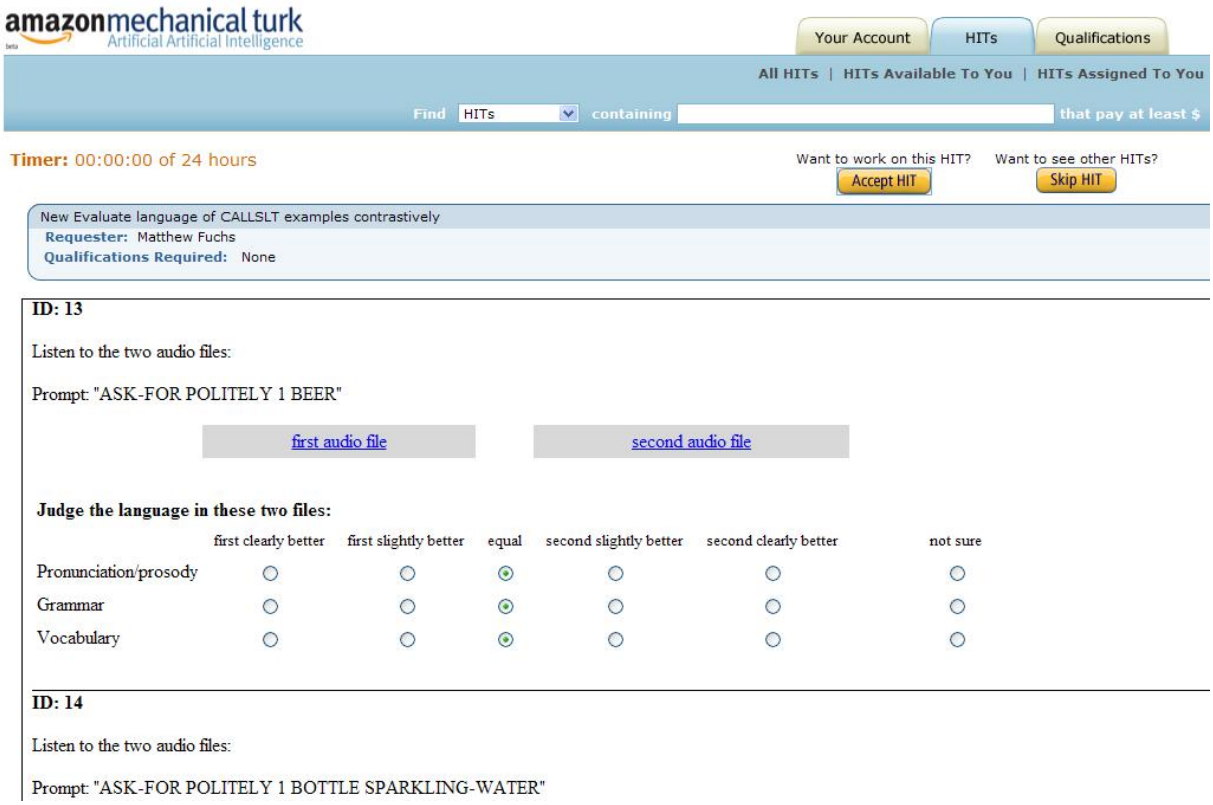

Figure 2: Amazon Mechanical Turk interface used for contrastive judging of recorded responses.

The notion of "quality of spoken response" is slippery; since we refrained from giving detailed guidelines, we were not surprised to see a fair degree of disagreement between the three judges. Even with respect to vocabulary and grammar, which one might expect to be reasonably uncontroversial, we found many differing judgements. For example, one judge thought a full sentence was grammatically better than a nominal phrase, while the other two considered them equally good. However, when we look at the "majority" judgements, we find a reassuring correlation between the human and mechanical evaluations; the judges agreed with the recogniser three times as often as they disagreed with it (90 versus 30). It is also worth noting that there are few cases of unanimous disagreements, and that, even when all the judges unanimously disagree with the recogniser, they often do not disagree for the same reasons: for example, one judge may think that the rejected utterance was better due to pronunciation, and another due to grammar.

6. Conclusions and further directions

It is notoriously difficult to evaluate CALL systems objectively (Chapelle, 2001; Chapelle, 2010), and the current experiment illustrates some typical problems. CALL-SLT implements a fairly ambitious strategy. It encourages students to formulate semi-free spoken responses to prompts, allowing them to improve their fluency and associated generative language skills, in contrast to more typical systems for pronunciation practice which require a specific response. Since the system does not know the response the student is trying to make, it is difficult for it to offer detailed advice on how they should try to improve their pronunciation; as with (Wang and Seneff, 2007) and subsequent systems, it only accepts or rejects.

When the student pronounces badly, the downside is that it may be unclear to them why the system rejected their utterance, particularly as they may feel that another utterance, which they pronounced less well, was accepted. To some extent, Table 2 shows that even experts can sometimes be confused in these cases. The problem is that beginner-level students hardly ever pronounce anything absolutely correctly. Given two imperfect pronunciations of the same utterance, the question of which one is "better" is to some extent subjective.

The upside is that, as Table 1 demonstrates, good pronunciation is usually accepted. A student who is skilful at imitating what they hear can improve their performance by listening carefully to the recorded native-speaker help examples and trying to move their own pronunciation closer to them; the key question is whether the coarse-grained accept/reject feedback given by the recogniser is helpful when the student is still some distance from their goal.

Our intuitive observation is that at least some students benefit from this kind of practice, though we do not feel that we can make strong claims yet; it is hard to construct tight experiments. In the study described here, we observed statistically significant improvements between the first and second halves of the session on at least some students (Bouillon et al., 2011). Unfortunately, given the difficulties with sound quality described in Section 3., it is challenging to separate linguistic improvement from simple acclimatisation to

Judge	Agree	Disagree	Null
All judgements			
1	82	40	48
2	87	31	52
3	99	51	20
at-least-1	134	80	7
majority	90	30	50
unanimous	44	12	114
Pronunciation and prosody judgements only			
1	72	39	59
2	76	21	73
3	93	45	32
at-least-1	127	69	11
majority	78	25	67
unanimous	36	11	123
Grammar and vocabulary judgements only			
1	59	18	73
2	50	22	98
3	41	20	109
at-least-1	84	43	57
majority	50	12	108
unanimous	16	5	149

Table 2: Agreement between system responses and human judgements on 170 well-recorded contrastive pairs. "Agree" means the judge(s) marked the element of the pair accepted by the system as better; "Disagree" means they marked it as worse; "Null" means no preference.

the peculiarities of the interface. A striking anecdotal result, though, derives from the fact that the "help" recordings used in the experiment had been recorded by Gerlach, who has an accent characteristic of the Vaud region of Switzerland; by the end of the session, a couple of the students (possibly more) had started to pronounce certain words using the same accent, which they were certainly not doing at the beginning. Another study (Rayner et al., 2011), using the Japanese-language version of CALL-SLT, provided unequivocal support for the claim that the system can help students acquire linguistic knowledge; it was however less clear that it helped them improve their pronunciation.

Looking ahead and attempting to learn from these experiences, some natural thoughts occur. A simple experiment, which we hope to perform in the near future, would be an extended form of the preliminary study reported in Table 1. By eliciting similar data from a wider range of subjects and asking native speakers for intuitive judgements on the relative linguistic abilities of the speakers, we would be able to determine whether these do indeed correlate with simple measures like WER and SER. It is not guaranteed that this would be the case; in the case of French, one could, for example, argue that, although speakers of closely related languages like Italian and Spanish are generally easy to understand and intuitively speak well, they tend to systematically mispronounce some common sounds, which could have a large effect on WER/SER. Similar remarks would apply to other languages which have close linguistic neighbours.

Another point arises from the lack of agreement between judges described in Section 4.. Our impression, based on this exercise, is that we would get considerably better agreement if we asked the judges more fine-grained questions; thus, for example, not "was utterance 1 pronounced better than utterance 2?" but rather, for example, "was the speaker's pronunciation of the following specific word/sound better in utterance 1 than in utterance 2?". This kind of approach has been suggested to us by some professional language teachers; it remains to be seen, however, whether it can feasibly be implemented in practice.

Ultimately, however, it is increasingly clear to us that the kind of small-scale study described here will not answer the key question: does the CALL system actually help students acquire useful language skills? This point is made at length in (Chapelle, 2001), and it is a good one. A truly convincing experiment needs to be carried out over a substantial length of time, and demonstrate that use of the tool results in skills that can be retained and carried over to real-world situations. It is evidently not easy to conduct a study which responds to these criteria, but it also seems to us that serious progress will be difficult unless such studies can be carried out.

Our long-range plan is to try to move in this direction. Meanwhile, we hope that small experiments, like this one, can at least throw light on some of the simpler aspects of the problem.

7. References

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

C. Chapelle. 2001. *Computer applications in second language acquisition: Foundations for teaching, testing and research*. Cambridge University Press.

C. Chapelle. 2010. Evaluating computer technology for language learning. *Contact*, 36(2):56–67.

H. Franco, H. Bratt, R. Rossier, V. Rao Gadde, E. Shriberg, V. Abrash, and K. Precoda. 2010. Eduspeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, 27(3):401.

W.L. Johnson and A. Valente. 2009. Tactical Language and Culture Training Systems: using AI to teach foreign languages and cultures. *AI Magazine*, 30(2):72.

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

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

M. Rayner, I. Frank, C. Chua, N. Tsourakis, and P. Bouillon. 2011. For a fistful of dollars: Using crowd-sourcing to evaluate a spoken language CALL application. In *Proceedings of the SLaTE Workshop*, Venice, Italy.

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