For A Fistful Of Dollars: Using Crowd-Sourcing To Evaluate A Spoken Language CALL Application

Manny Rayner[1], *Ian Frank*[2], *Cathy Chua*[3], *Nikos Tsourakis*[1], *Pierrette Bouillon*[1]

[1] University of Geneva, ETI/TIM/ISSCO, Geneva, Switzerland
[2] Future University Hakodate, Japan
[3] Faculty of Information and Communication Technologies,
Swinburne University of Technology, Melbourne, Australia
{Emmanuel.Rayner,Nikolaos.Tsourakis,Pierrette.Bouillon}@unige.ch,
ianf@fun.ac.jp, cathychua@swin.edu.au

Abstract

We present an evaluation of a Web-deployed spoken language CALL system, carried out using crowd-sourcing methods. The system, "Survival Japanese", is a crash course in tourist Japanese implemented within the platform CALL-SLT. The evaluation was carried out over one week using the Amazon Mechanical Turk. Although we found a high proportion of attempted scammers, there was a core of 23 subjects who used the system in a responsible manner. The evidence that these subjects learned from their 111 sessions and 9092 spoken interactions was significant at $P = 0.001$. Our conclusion is that crowd-sourcing is a potentially valid method for evaluating spoken CALL systems.

Index Terms: CALL, crowd-sourcing, speech recognition, evaluation, Japanese

1. Introduction

Evaluation of CALL systems is often problematic. The most common approach, at least in our experience, is to round up a bunch of students, sit them down in front of a few laptops, persuade them to use the system, and record how they got on. This has several advantages. It is possible to keep tight control over the experiment, and to check that the students are doing what is expected of them; also, one can often arrange to get a more or less uniform sample, typically a group of people with similar ages and educational backgrounds. There are also several related disadvantages. Having a uniform sample may not necessarily be a good thing, since it leaves open the possibility that the system is too closely tuned to that type of student. Another problem is that logistical considerations often make it difficult to get students to use the system for multiple sessions, so that one can track their progress.

In this paper, we describe an experiment using CALL-SLT [1], a spoken language CALL system for practising fluency in a limited domain based on the "spoken translation game" idea of Wang and Seneff [2]. We have previously carried out a number of evaluations of CALL-SLT using different versions of the "round up some students" approach sketched above [1, 3, 4]. Here, we tried a different approach, and used crowd-sourcing methods to recruit subjects through the Amazon Mechanical Turk (AMT). These subjects were asked to test a version of the system loaded with an introductory Japanese course ("Survival Japanese") suitable for people who wished to spend a few hours acquiring a smattering of basic words, phrases and grammar before visiting the country. The course was designed to be accessible to people who had no previous exposure to the language. The content of the course seemed to be a reasonable fit to crowd-sourcing methods.

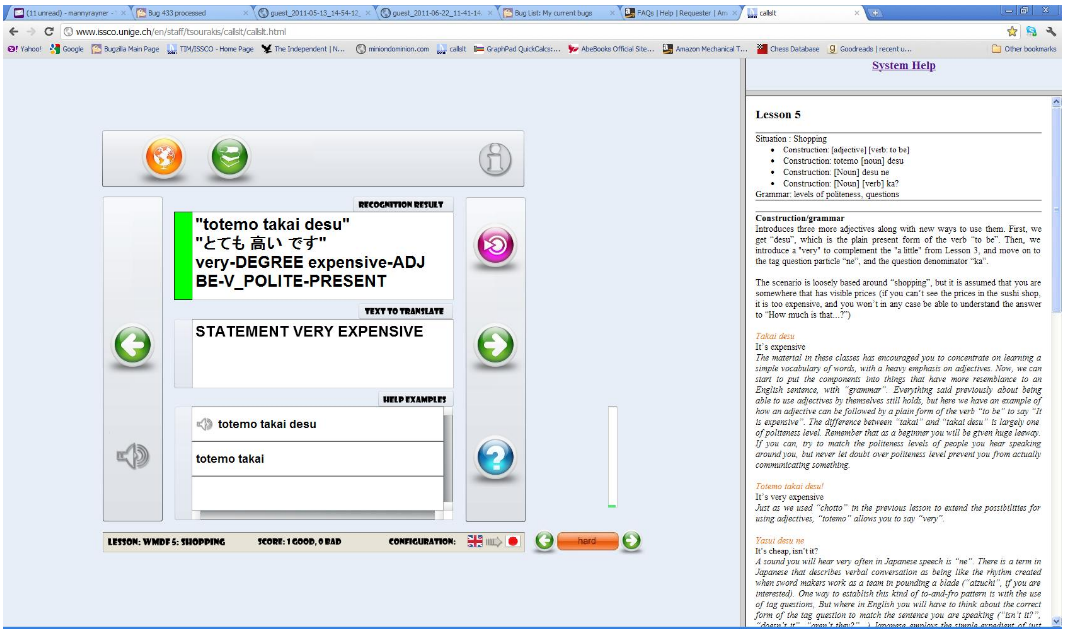
We had two goals in mind: we wanted to evaluate both the CALL system and the feasibility of the data collection methodology. Over the last couple of years, the use of crowd-sourcing for the collection of data for Spoken Dialogue Systems has become increasingly topical [5]. Conventionally these projects have followed the KISS principle, which may be appropriate for the building of speech databases [6], but we wondered if crowd-sourcing could be used more ambitiously for the qualitative evaluation of a CALL system, a difficult task. Far from wanting workers to do the same ten second job over and over again, by repute the way to get best results from a "labourforce" that is not highly regarded for the level of its capabilities, we wanted subjects to do no less than learn to speak a language, and that over a few sessions over a few days. The task was sophisticated and, by crowd-sourcing standards, time-consuming for the subject. This gave us an interesting opportunity to test the common belief that such tasks are inappropriate for crowd-sourcing due to the poor quality of the workers.

The rest of the paper is organised as follows. Sections 2 and 3 give background on the CALL-SLT system and the Survival Japanese course. Section 4 presents the experiments themselves, and Section 5 the results. The final section concludes.

2. The CALL-SLT System

CALL-SLT is an Open Source speech-based CALL application for beginning to intermediate-level language students who wish to improve their spoken fluency. It leverages earlier work on Regulus, a platform for building systems based on grammar-based speech understanding [7] and MedSLT, an interlingua-based speech translation framework [8], to develop a generic CALL platform centered on the "spoken translation game" idea. The system is deployed over the Web using a server/client configuration. Most processing, in particular speech recognition and language understanding, occurs on a remote server. The client is a Flash process running inside a normal web-browser. Figure 1 presents a screen-shot.

Our experiences to date suggest that the Regulus/MedSLT architecture is a good fit to this type of application. In particular, the grammar-based approach to recognition gives a response profile with accurate recognition on in-grammar utterances and



RECOGNITION RESULT

"totemo takai desu"
"とても 高い です"
very-DEGREE expensive-ADJ
BE-V_POLITE-PRESENT

TEXT TO TRANSLATE

STATEMENT VERY EXPENSIVE

HELP EXAMPLES

totemo takai desu

totemo takai

LESSON: WMDF 5: SHOPPING SCORE: 1 GOOD, 0 BAD CONFIGURATION:

System Help

Lesson 5

Situation : Shopping
- Construction: [adjective] [verb: to be]
- Construction: totemo [noun] desu
- Construction: [Noun] desu ne
- Construction: [Noun] [verb] ka?

Grammar: levels of politeness, questions

Construction/grammar
Introduces three more adjectives along with new ways to use them. First, we get "desu", which is the plain present form of the verb "to be". Then, we introduce a "very" to complement the "a little" from Lesson 3, and move on to the tag question particle "ne", and the question denominator "ka".

The scenario is loosely based around "shopping", but it is assumed that you are somewhere that has visible prices (if you can't see the prices in the sushi shop, it is too expensive, and you won't in any case be able to understand the answer to "How much is that...?")

Takai desu
It's expensive
The material in these classes has encouraged you to concentrate on learning a simple vocabulary of words, with a heavy emphasis on adjectives. Now, we can start to put the components into things that have more resemblance to an English sentence, with "grammar". Everything said previously about being able to use adjectives by themselves still holds, but here we have an example of how an adjective can be followed by a plain form of the verb "to be" to say "It is expensive". The difference between "takai" and "takai desu" is largely one of politeness level. Remember that as a beginner you will be given huge leeway. If you can, try to match the politeness levels of people you hear speaking around you, but never let doubt over politeness level prevent you from actually communicating something.

Totemo takai desu!
It's very expensive
Just as we used "chotto" in the previous lesson to extend the possibilities for using adjectives, "totemo" allows you to say "very".

Yasui desu ne
It's cheap, isn't it?
A sound you will hear very often in Japanese speech is "ne". There is a term in Japanese that describes verbal conversation as being like the rhythm created when sword makers work as a team in pounding a blade ("aizuchi", if you are interested). One way to establish this kind of to-and-fro pattern is with the use of tag questions, But where in English you will have to think about the correct form of the tag question to match the sentence you are speaking ("isn't it?",

Figure 1: Screenshot of a CALL-SLT Survival Japanese lesson running as a client in a web-browser window. The left pane contains the prompt (in this case, "statement very expensive"), with user-requested help examples below and any recognition result above (presented in alphabetic characters, native script, and as a gloss). The right pane shows the detailed lesson explanation.

poor or no recognition on out-of-grammar utterances, automatically giving the student feedback on the correctness of their language usage. Also, the platform's rapid development facilities, based on semi-automatic specialisation of general resource grammars, have made it easy to create multiple good speech recognisers. Although the recognisers for the L2 languages are all built from development corpora of at most a few hundred examples, native speakers typically get per-sentence semantic error rates of under 10%.

One way that CALL-SLT differs significantly from Wang's and Seneff's work is in its presentation of prompts to the students. Instead of giving students target sentences in their own language (the L1), our system uses interlingua representations, which are created using semantic grammars based on our previous work on human-readable representations of interlingua [8]. By prompting using interlingua forms, it is possible to reduce the undesirable effect of tying the language being studied (the L2) too closely to the L1 in the student's mind, as recommended by mainstream theories of language acquisition. In the experiments described in this paper, the interlingua is realised in a telegraphic textual form based on English (we also support forms based on other L1s, including French, Japanese, Chinese and Arabic), so the result is still relatively close to the straightforward L1 prompts that Wang and Seneff employ. It is possible, however, to produce graphical realisations of the interlingua [1] without changing the underlying architecture.

In more detail, the game that forms the basis of CALL-SLT is as follows. The system is loaded with a set of possible prompts that represent the target content for a given lesson. Each turn starts with the student asking for the next prompt. The system responds by showing a surface representation of the underlying interlingua for a target L2 sentence. For example, a student whose L1 is French and whose L2 is English might be given the following textual prompt:

```
COMMANDER DE_MANIERE_POLIE SOUPE
```

An appropriate response would be something like "Could I have the soup?", "I would like some soup", or simply "Soup, please"; the grammar supports most of the normal ways to formulate this type of request.

After deciding what to say, the student presses the "recognise" button, and speaks. The system performs speech recognition using a Nuance Toolkit recognition package compiled from a grammar-based language model, 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. If the match was successful and the student is registered as a native speaker, their recorded speech is also saved for future use.

The student may ask for help at any time. The system can give help either in speech or text form. Text help examples are taken from the original corpus, and can also be produced by translating from the interlingua back into the L1; speech help examples are created by recording successful interactions.

The prototype CALL-SLT system is freely available for use; see `callslt.org` for detailed instructions. Currently, the system offers about a dozen combinations of L1s and L2s.

3. The Survival Japanese Course

The Survival Japanese course was designed by Ian Frank, a native English speaker with Japanese language fluency. The content focuses on extremely simple communication with an emphasis on adjectives. The goal is to enable students to quickly reach a level where they feel they can be part of the group in any social situation by always being able to contribute something.

Japanese is a prodrop language notorious for omitting surface elements that can be inferred from context. Adjectives are closely related to verbs, and verbs do not require a subject. A lone adjective is thus a well-formed clause, which can be modified by attitude and degree particles. So, for example, while an English speaker might say “It’s hot!” or a French speaker *Il fait chaud!*, in Japanese it is enough to say *Atsui!* This can be turned into a tag question by adding the particle *ne*; thus, *Atsui ne?* means “It’s hot, isn’t it?”. A few patterns like these are enough to provide a surprising range of possibilities.

The version of the course used here introduced 16 adjectives. Since we are collaborating with a music festival in Japan (WMDF; `wmdf.org`) the lessons are loosely themed around a short artist tour, but we tried to make the materials widely applicable (see Table 1). In general, the course is designed to progress incrementally by presenting language in the simplest possible (shortest) way first, before review and expansion in later lessons. Recently, a group of eight artists (native English speakers with extremely limited Japanese knowledge) visited WMDF for one week and used a printed summary version as a handy reference which they carried with them. All members reported using “80% or more” of the language at least once (half said they used everything) and three quarters described the level of language content as “About right” with one quarter replying “Complicated but understandable”. Their enthusiastic responses encourage us to believe the course content was useful for its purpose.

Theme	#	Topics	Example language
Airport	7	First adjectives, Simple phrases	*Hajimemashite* (Nice to meet you)
Greetings	5	Getting by when meeting people	*Sumimasen* (Excuse me/I’m sorry)
Restaurant	8	Requesting, More adjectives	*Chotto onegai shimasu* (A little, please)
Stage	8	Adjectives again, More on *chotto*	*Sugoi!* (Great!)
Shopping	9	*ne* tag particle, Yes/no	*Yasui desu ne?* (It’s cheap, isn’t it?)
Adjective use drill	21	Review of ten sentence patterns	*Chotto atsui desu ka?* (Is it a bit too hot?)
Party	15	Polite forms, Grammar check	*Tanoshii deshou ka?* (Are you having fun?)
Farewell	9	Past tense (verbs and adjectives)	*Subarashikatta desu* (It was wonderful)

Table 1: Summary of Survival Japanese course: lesson theme, number of examples (#), topics, and example target language.

4. Experimental Setup

The unit of work on AMT is the “Human Intelligence Task” (HIT). We recruited subjects to work on our HITs in two ways:

1. Inviting AMT workers from a previous experiment with a speech-enabled internet game (the experiment was an extended version of the one described in [9]). These subjects had previously scored at least 90% on “HIT acceptance rate” (i.e., at least 90% of their previous tasks had been accepted), and they had also performed reliably in our experiment. These subjects were all US residents.
2. A general advertisement on AMT offering a small amount of money to go to our web site, try out the system and register to express interest. Constraints on HIT acceptance rate and location were not imposed.

Subjects from the first group again performed responsibly, but those from the second were extremely disappointing: all but one turned out to be scammers. It is clear that for the sort of HITs we wish to post, some form of preselection of workers significantly reduces scamming. The internet game experiment referred to above agrees with this observation.

As in our earlier experiment, we quickly discovered that we needed to pay more than the $1/session we started out offering. We increased to $2/session with $5 for the last session, in which we expected more, including a test without using the help function. All subjects were given a unique ID, so that we could associate session logs with information returned through AMT. We told subjects that, over a period of a few days, they could do up to 8 sessions of 10–20 minutes duration, speaking at least 30 utterances each time.

Beyond that we did not impose any particular methodology. This seemed appropriate to the nature of the course, which was designed to be done individually and in the user’s preferred way. We did not want to put subjects off by making the process too formal or inflexible and the Survival Japanese course itself was designed to encourage the user to put in effort to advance further. Although we could track the subjects’ progress via the session logs, we asked for feedback throughout as we also wanted their impressions of how they were doing. We wished not only to get an idea of whether the system was a psychological boost to language learning, increasing the enjoyment factor by making something more akin to a game, but also to see in what way subjects’ impressions of how they were going related to the actual log data of their performance during the trial.

We did, however, attempt to nudge subjects in the direction of more systematic use of the system, by including questions in the AMT feedback forms that encouraged them at least to think about what their learning strategy was, and what progress they were making on the individual lessons. Several subjects answered early on that they intended to practise carefully and master the course. As we show in the next section, many made good on their promises.

5. Results

We posted seven HITs on AMT over a one week period, releasing approximately one HIT per day. A total of 130 workers responded to at least one of the HITs. Of these, we found that a surprising total of 94 were lying, and had never logged in. The majority did not even have a login ID and password.

There were 26 subjects who logged in successfully and left data on the server. After examining summaries of the logfiles, we found that we could separate them into three groups, which we dubbed “serious”, “casual” and “scammers”. Except for two or three borderline cases, it was in general quite clear which group a subject belonged to.

Scammers were, again, flat-out lying, but had been organised enough to get as far as interacting with the system. They performed few or no interactions, got nothing recognised correctly, had very short sessions, and left deceptive comments. We found 3 clear scammers. In general, scammers were detected by a few simple scripts that analysed the session logs left on the server. They were not paid, except for a few who sneaked past at the beginning before the scripts were fully functional.

“Casuals” were people who came in, genuinely interacted with the system, but did not appear to have any serious intention of learning anything. Most of them just seemed to think it was fun to experiment, and then left, though a few came back for one or two repeat sessions. In total, we had 11 “casuals”, who together did 29 sessions.

"Serious" subjects came back multiple times, and gave strong evidence of trying to learn the course. They practised the lessons systematically, in most cases showed measurable improvements, and left insightful comments. We had 12 "serious" subjects, who together did 82 sessions.

According to the AMT trace files, the non-scammers together logged a total of 36 hours of interaction with the system for a cost of $170, averaging 22 minutes and $4.74 per session. It is striking that, although we had asked for 10–20 minutes in the instructions, the *average* session length was greater than our suggested upper limit.

Type	Subj	Se	Rec	R/Se	✓%	Help%
"serious"	12	82	7734	94.3	34.5	60.4
"casual"	11	29	1358	46.8	34.2	59.7
"scammer"	3	18	18	1.0	0	51.3

Table 2: Summary data for "serious", "casual" and "scammer" subjects: number of subjects in each group, number of sessions, number of recognition events, recognition events per session, proportion of correct matches (✓) on recognition events, and proportion of times help was accessed.

Table 2 summarises the data for the three types of subject. The "serious" and "casual" subjects received almost exactly the same scores in terms of recognition and using help[1]. The main difference, however, does not appear to be a question of ability, but rather of motivation; the "serious" subjects not only averaged many more sessions, but also had many more recognition events per session.

Round	Subj	Prompts	Rec	✓%	Help%
2	12	880	1971	31.4	67.4
3	10	653	1119	37.0	60.9
4	10	525	839	34.9	49.0
5	7	265	592	32.9	50.2
6	7	240	418	37.6	29.2

Table 3: Summary data for "serious" subjects, rounds 2 to 6: round number, subjects remaining, number of recognition events, proportion of correct matches (✓) on recognition events, and proportion of prompts for which subjects accessed help.

Nearly all the "serious" subjects left comments suggesting that they believed that they were learning something, and we sought objective evidence to support this claim. Some straightforward data is presented in Table 3, which tracks the progress of the "serious" subjects over the HITs we posted from rounds 2 to 6. (The data from round 1 was unfortunately rendered invalid by a log-file bug, and only four subjects did round 7). Recognition performance did not change much, oscillating unevenly between 31% and 37%. Average help usage, however, decreased a great deal, dropping smoothly from 67.4% in round 2 to 29.2%; this difference is significant at $P = 0.001$ according to the Fisher test (two-tailed). The obvious interpretation is that subjects were retaining more and more vocabulary and grammar, just as they claimed, and by the end were often able to remember things for themselves without looking them up. It was easy to eliminate the other main hypothesis, namely that the improvement was due to the less skillful subjects dropping out: in fact, average help usage improved slightly more when we only considered the seven subjects who reached round 6.

6. Summary and Conclusions

As previously mentioned, we had two goals in mind when carrying out this experiment. First, we wanted to evaluate Survival Japanese: was it possible to use CALL-SLT to build a basic spoken language course for beginners that could be done over a few days, in odd moments, but would still be tangibly useful to people who were prepared to put in a little effort? We still need to perform a more fine-grained analysis of the logged data, but the gross figures in Table 3 suggest that the course is, at the very least, a partial success.

Our top priority, however, was to investigate whether crowd-sourcing is a valid methodology for evaluating spoken language CALL systems. On balance, our feeling is guardedly positive. It has been rewarding to work with the group we called the "serious" subjects. These people did more than was asked of them, subjecting CALL-SLT and Survival Japanese to a rigorous test that revealed many important problems in our system. Our impression, based on this initial attempt, is that the "serious" profile is not too hard to identify. It is straightforward to write scripts that unmask the scammers. Distinguishing "serious" from "casual" subjects is more challenging, but this is not a critical problem; since the "casual" subjects, by definition, disappear quickly, most of the sessions end up being done by the "serious" subjects who stay in. We plan to run another evaluation soon, where we will test these hypotheses empirically and attempt to recruit a larger subject pool.

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] P. Bouillon, I. Halimi, M. Rayner, and N. Tsourakis, "Evaluating A Web-Based Spoken Language Translation Game For Learning Domain Language," in *Proceedings of INTED 2011*, Valencia, Spain, 2011.

[4] P. Bouillon, M. Rayner, N. Tsourakis, and Q. Zhang, "A Student-Centred Evaluation of a Web-Based Spoken Translation Game," in *Proceedings of this conference*, 2011.

[5] C. Callison-Burch and M. Dredze, Eds., *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, CA, 2010.

[6] I. McGraw, C. Lee, L. Hetherington, S. Seneff, and J. Glass, "Collecting voices from the cloud," in *Proc. LREC*, 2010.

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

[8] P. Bouillon, G. Flores, M. Georgescul, S. Halimi, B. Hockey, H. Isahara, K. Kanzaki, Y. Nakao, M. Rayner, M. Santaholma, M. Starlander, and N. Tsourakis, "Many-to-many multilingual medical speech translation on a PDA," in *Proceedings of The Eighth Conference of the Association for Machine Translation in the Americas*, Waikiki, Hawaii, 2008.

[9] C. Chua and M. Rayner, "What's the magic word?" in *Proc. Thirteenth Australasian International Conference on Speech Science and Technology*, Melbourne, Australia, 2010.

[1] The "serious" group arguably performed better, since they did a slightly higher proportion of examples from the harder lessons.