Using a Serious Game to Collect a Child Learner Speech Corpus

Claudia Baur, Manny Rayner, Nikos Tsourakis

Université de Genève FTI / TIM
40, bvd Du Pont-d'Arve, CH-1211 Genève 4, Switzerland
Claudia.Baur@unige.ch, Emmanuel.Rayner@unige.ch, Nikolaos.Tsourakis@unige.ch

Abstract

We present an English-L2 child learner speech corpus, produced by Swiss German-L1 students in their third year of learning English, which is currently in the process of being collected. The collection method uses a web-enabled multimodal language game implemented using the CALL-SLT platform, in which subjects hold prompted conversations with an animated agent. Prompts consist of a short animated Engligh-language video clip together with a German-language piece of text indicating the semantic content of the requested response. Grammar-based speech understanding is used to decide whether responses are accepted or rejected, and dialogue flow is controlled using a simple XML-based scripting language; the scripts are written to allow multiple dialogue paths, the choice being made randomly. The system is gamified using a score-and-badge framework with four levels of badges. We describe the application, the data collection and annotation procedures, and the initial tranche of data. The full corpus, when complete, should contain at least 5 000 annotated utterances.

Keywords: child speech, corpora, CALL, serious games

1. Introduction

The process of collecting spoken corpus data has traditionally been tedious and expensive, but recent technological advances have opened interesting new possibilities. In particular, the fact that speech-enabled applications can now easily be deployed on the internet makes it possible to distribute data-collection work far more efficiently than was previously the case; this idea can often be combined with crowd-sourcing, using sites like the Amazon Mechanical Turk to send requests to a large pool of prospective users and automate the process of harvesting data and handling payments (McGraw et al., 2010; Jurčiček et al., 2011).

In the present paper, we describe how these techniques can be adapted to the task of collecting a corpus of child learner speech, specifically speech by German-speaking children learning English. We created an online language-learning game designed for beginner German-speaking students, and deployed it on the web. The game, implemented using the CALL-SLT platform (Rayner et al., 2010), uses speech recognition to give users an interactive learning experience and encourage them to produce spontaneous speech. Data is automatically logged in an easily manipulable tabular form, which can then be efficiently cleaned up by human annotators. We present a concrete data collection exercise, using students at a school in German-speaking Switzerland who are currently in their third year of learning English.

In the rest of this paper, we first describe the language-learning game and then the data collection and annotation process with the achieved results.

2. The Language Game

2.1. Basic Functionality and Architecture

CALL-SLT is a prompt-response system based on speech recognition and machine translation technology, in which the system and the user take alternate turns; CALL applications of this general kind can be traced back to the "spoken translation game" described in (Wang and Seneff, 2007). Each interchange begins with the system giving the student a prompt, which in the present version is a combination of an English-language multimedia file and a piece of German text; for example, at the beginning of the Hotel lesson, the system plays a cartoon clip of a desk clerk asking "How many nights would you like to stay at our hotel?", simultaneously showing the German text "Frag: Zimmer für 6 Nächte" ("ask: room for 6 nights"). The student then gives a spoken response; the intent is that this should be reasonably free, so here one can for instance answer "I would like to stay for six nights", "I want a room for six nights", "A room for six nights please", and several other variants.

The system decides whether to accept or reject the response by first performing speech recognition, then translating to a language-neutral (interlingual) representation, and finally matching this representation 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; spoken help examples are collected from previous successful interactions. When the system has decided whether to accept or reject, it moves to a new dialogue state; the choice of state is determined by an XML-based script written by the course designer, which specifies various options. Continuing the example, an "accept" moves to a state where the desk clerk's next question is "What type of room would you like?"; a "reject" stays in the same state, with the desk clerk saying that he didn't understand; and a sequence of three rejects moves to a state where the clerk says he didn't understand, but asks whether a room for one night will be okay. On reaching the end of the lesson, the student either exits or selects a new lesson from a menu. Figure 1 shows the user interface.

We briefly summarize the underlying architecture, which has been described at length in earlier papers (Rayner et al., 2010; Rayner and Tsourakis, 2013; Baur et al., 2013). Speech and language processing use a grammar-based framework embodied in the Regulus platform (Rayner et

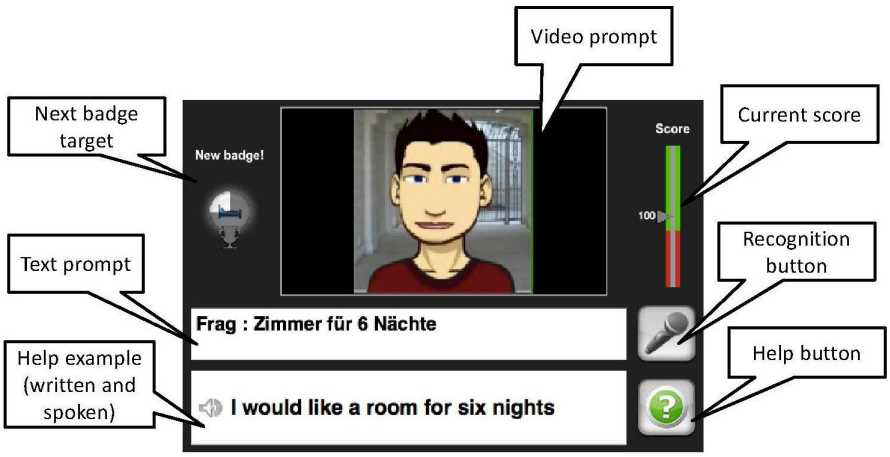

Figure 1: Screenshot of CALL-SLT user interface.

al., 2006), which in turn sits on top of the commercial Nuance Toolkit package. The core idea is reusability. A general resource grammar, written in a feature-grammar notation, is defined for the recognition language (here, English) and shared between applications. Efficient domain-specific feature grammars are compiled out of it using a version of the Explanation Based Learning algorithm, driven by small corpora of examples; these grammars are then further compiled into the proprietary CFG form required by the Nuance Toolkit. The Nuance Toolkit performs two more compilation steps, first using the example corpus to add probabilistic weights to the CFG form, and then converting the resulting PCFG grammar into a Nuance language model. In practice, the training corpus is usually partitioned into smaller subcorpora, with each subcorpus producing a different language model. This makes it easy to construct tightly constrained recognizers which offer good performance on in-coverage utterances.

In the context of the CALL-SLT application, the corpus used for training the recogniser also defines the examples used for each lesson. We construct a language model for each individual lesson, covering the examples listed for that lesson, and various generalizations of them obtained by combining rules and lexical entries taken from different training examples. We have also experimented with compiling models for unions of sets of lessons, for example for the union of all the lessons, or for the current lesson together with all the preceding lessons. The tradeoff with these larger models is between increased coverage and degraded recognition. For beginner students like the ones we are dealing with here, forgiving recognition seemed more important, and we consequently used only the small lesson-specific language models.

The same specialised grammars used to construct the language models are also used to parse recognition results: every recognition result is thus grammatically well-formed. Semantic representations are produced in a minimal feature-value notation called Almost Flat Functional Semantics (AFF; (Rayner et al., 2008). For example, "Could you give me directions to the zoo?" is represented as the structure

```
[null=[utterance_type, ynq],
 agent=[pronoun, you],
 null=[modal, could],
 null=[action, give],
 null=[voice, active],
 object=[abstract, directions],
 indobj=[pronoun, i],
 to_loc=[loc, zoo]]
```

L2 – here, English-language – AFF representations are mapped into language-neutral (interlingual) counterparts, using a set of rules which rewrite tagged lists of elements to tagged lists of elements. This is done in a way which tries to map semantically equivalent L2 utterances to the same interlingual utterance, so that matching between prompt and response can be performed at the interlingual level. A second grammar is responsible for realizing the interlingual forms in the L1, here German. Thus, to continue the example, the English AFF form above is mapped into the interlingual AFF form

```
[null=[utterance_type, request]
 arg2=[abstract, directions],
 to=[loc, zoo]]
```

which is realized in German as the prompt "Frag : Wo ist der Zoo ?" ("Ask: where is the zoo?"). The same representation is produced from e.g. "Where is the zoo?", "I am looking for the zoo?" and "Can you tell me where the zoo is?".

The system is deployed over the web using a scalable architecture designed for cloud-based computing. In common with similar platforms, like WAMI (Gruenstein et al., 2008) and Nuance's Mobile Developer Platform, it uses a client/server approach in which speech recognition is carried out on the server side. The overhead due to web deployment is a few hundred milliseconds per recognition operation, compared to execution on a desktop machine. Full details are presented in (Fuchs et al., 2012).

2.2. Content

The course content was formulated in close collaboration with an English teacher working at a secondary school in

German-speaking Switzerland, who provided the subject matter expertise required to guarantee the usefulness of the content for the target audience. On the basis of this collaboration, we decided to link the content to the the school's curriculum and the standard English textbook *Ready For English 1* used in Swiss German schools (Morrissey et al., 2001), as well as to use a communicative approach to second language learning, including multimedia elements to simulate a conversation partner for the language learner. This resulted in a dialogue-based system, loosely covering the first year's class content. We created the following eight coherent lessons or dialogues, which can be put together to simulate a virtual trip to London:

Train station: name, nationality, numbers, locations, time expressions

Introducing yourself: name, nationality, siblings, capitals

Tube station: numbers, locations, prices

Hotel: numbers, room types, prices, payment types, where-questions

Tourist Information Office: numbers, cultural knowledge of London, time expressions, ordinal numbers

Restaurant: food and beverages, payment types

Asking and giving directions: where-questions, directions, distances

Shopping: clothing, colours, numbers, like/dislike expressions.

This approach allows the students to practice their receptive skills (by listening to the videos recorded by native English speakers), as well as their productive skills (by engaging in the conversation and responding to the questions), two closely linked components of second language acquisition according to the CEFR (Council of Europe, 2001). Table 1 shows examples of typical content.

The eight lessons use a combined vocabulary of about 450 words. A dialogue typically contains between 10 and 20 states (average of 14.4); in each state, there are typically between 5 and 15 possible prompts (average of 10.8) that can be issued. Random choices are made both for prompt and state transitions, so students get a different interaction each time they do a lesson. The dialogue flow is structured so that students are given at most two attempts at each prompt, after which the system backs off to a yes/no question: the intention is to make sure that the student cannot get stuck at a difficult step.

In order to increase student's motivation and engagement in the course, we added some very basic gamification elements (Werbach and Hunter, 2012). The students lose points each time the system fails to recognize them, and gain points for using harder constructions like *could you ...* or *do you have ...*. This scoring system is directly linked to a badge system with four levels of badges. For the two lower levels (plain and bronze), the student only has to complete the lesson a specified number of times; at the higher levels (silver and gold), where the help function is also switched off, they need to complete the lesson enough times while also achieving a specified minimum score (85 and 90 points respectively; the student starts with 100). After every successfully completed lesson, the student has the choice between doing the same lesson again and collecting another badge or changing the lesson.

Train station
I would like two tickets to London
I need to leave on Tuesday afternoon
Introducing yourself
I am from Italy
I have two sisters
Tube station
Can I pay by credit card?
Thank you
Hotel
I would like to stay for three nights
Is there a swimming pool?
Tourist Information Office
I would like tickets for Wicked
I want to sit in the third row
Restaurant
I want my steak well done
I did not order this
Asking and giving directions
Where is the Science Museum?
Which bus goes there?
Shopping
I am a small
This is too expensive

Table 1: Examples of typical correct responses for the different lessons

Again in the interests of making the dialogue flow more interesting and increasing motivation, some lessons include extra subdialogues which are only activated when the student has advanced to the silver and gold levels. The unifying idea is "non-cooperative dialogue"; so, for example, in the 'Restaurant' lesson, the dialogue flow at the lower levels follows a simple path where the student is told to order a main course, something to drink, and a dessert, after which they pay and leave. At the higher levels, new options are added so that the waiter may bring the wrong order, after which the student needs to complain and if necessary ask to see the manager. We are currently in the middle of developing this idea, and will describe it in more detail elsewhere.

3. Data Collection and Annotation

3.1. Data Collection Framework

Data is automatically logged by the system, with a new timestamped logfile produced for each session. The logfile contains records of all the primitive operations carried out during the session, including logging in, starting a new lesson, being given a prompt, asking for help, producing a spoken response and being rated on the response. Recorded speech files, in SPHERE-headed form, are attached to the appropriate records. The transcription in the

SPHERE header is initially the result produced by speech recognition.
The raw logfiles are processed by a script which collects the data and turns it into HTML-formatted spreadsheets, one line for each prompt issued; this also creates copies of the recorded files in RIFF format, which is easier for most people to play. The RIFF files are copied to a webserver, so that the audio files linked from the spreadsheet are accessible through a normal browser. Each line of the spreadsheet contains the following information; the last five fields are initially left blank, and are later filled in by a human annotator.

1. Subject ID
2. Prompt (in general, a combination of a multimodal file ID and a piece of text)
3. Link to recorded file in RIFF format
4. Transcription
5. Whether help was accessed (yes/no)
6. Whether the student's response was accepted (yes/no)
7. Estimate of recording quality
8. Estimate of lexical quality
9. Estimate of grammatical quality
10. Estimate of phonetic quality
11. Estimate of fluency quality

The HTML speadsheets are annotated using Microsoft Excel. For each line, the annotator clicks on the link to listen to the recorded file. They correct the transcription if necessary and fill in the last five fields, using a 5-point scale where 1 is "completely wrong" and 5 is "near-native-speaker quality". Table 2 gives more details on the annotation labels.

3.2. Current Data Collection Exercise

Using the framework described in 3.1., we have collected data in collaboration with a school in German-speaking Switzerland, between October and November, 2013. A class of 19 students, who are a couple of months into their third year of English, are being used as subjects. They were encouraged to use the system as a supplement to their regular classroom instruction; it is for this reason that we have designed content which is consistent with their English textbook, as described in section 2.2..
The pupils were in a third year level P class (most advanced of three levels). Out of 19 students we had a total of 11 actual users, meaning that 58% of the students in the class used the system successfully. 7 (= 64%) were female and 4 (= 36%) were male users, with an age range between 13 and 16 years, however most users (64%) were 14 years old. The total of 862 logged interactions (all users together) was collected during class-time, where the students used the system in the school's computer lab and at home, where the students interacted with CALL-SLT by themselves as a homework assignment.
Two annotators – one English native speaker and one German native speaker with near-native English skills – corrected the transcriptions and filled in the empty fields in the spreadsheets.

Category	Possible labels
Recording quality	discard
	cut off
	high background noise
	low background noise
	high volume
	low volume
	non-linguistic interruption
Vocabulary	correct
	incorrect
	incorrect response
Grammar	correct
	incorrect
	incomplete
	incorrect response
Pronunciation	1 - incomprehensible
	2 - clear mistakes
	3 - correct
	4 - near native
	5 - native-like
Fluency	1 - incomprehensible
	2 - clearly unfluent
	3 - correct
	4 - near native
	5 - native-like
Fluency additional info	hesitation
	stressing
	intonation
	over-articulation

Table 2: Annotation labels

3.3. Annotation Results

From the total of 862 utterances, 48 had to be discarded (5.6%), since some files were empty and for others it was impossible to determine what the user was saying. This resulted in a total of 814 analysed and annotated utterances for the current sub-corpus.
Most of the utterances are spontaneous speech, recorded while interacting with the language game as explained in section 2.; there was also a small amount of read speech included (66 utterances; 7.7%). These data were collected in the "pre-test" lesson, which served as a placement test in the beginning of the data collection exercise. Read speech utterances are typically longer than spontaneous speech and count between 7 and 19 words (average of 15.3 words). As can be seen in Table 3, the length of spontaneous speech utterances is much shorter, with almost a quarter of all utterances (23%) consisting of as little as one word. Utterances with up to six words are common, but the frequency drops sharply after eight words.

Length [words]	Occurrences	Proportion
1	173	23.10%
2	68	9.08%
3	28	3.74%
4	114	15.22%
5	110	14.69%
6	110	14.69%
7	48	6.41%
8	62	8.28%
9	24	3.20%
10	8	1.07%
11	1	0.13%
12	1	0.13%
13	1	0.13%
14	1	0.13%

Table 3: Utterance length

Table 4 shows the measured weighted Cohen's Kappa (squared weights) for the ratings by the two annotators (using 95% C.I.), indicating a moderate agreement on the categories "vocabulary", "grammar", "pronunciation" and "fluency".

Category	Kappa	C.I.
Vocabulary	0.46	(0.32, 0.58)
Grammar	0.45	(0.32, 0.57)
Pronunciation	0.66	(0.61, 0.70)
Fluency	0.53	(0.48, 0.58)

Table 4: Cohen's Kappa

The overall annotation distribution for the categories "pronunciation" and "fluency" by the two annotators, divided into spontaneous and read speech, are displayed in Figures 2 and 3. Figures 4 and 5 indicate the distribution of correct and incorrect utterances.

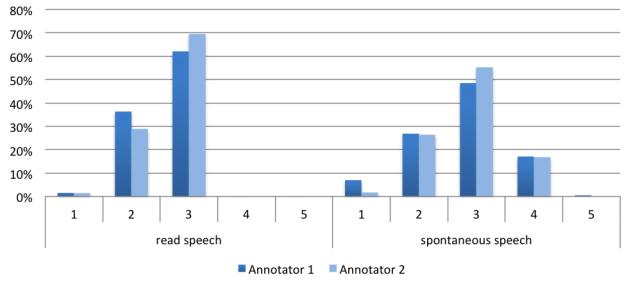

Figure 2: Pronunciation results read and spontaneous speech

Inter-annotator agreement on recording quality was unfortunately less good. Figure 6 shows that annotator 1 judged more than half of all utterances (56%) to have a low level of background noise and 12% as having high background noise. The number of recordings with very high or low recording volume (0.6% and 2.6% respectively) can be neglected for this annotator. Annotator 2 judged less utterances as having (high or low) background noise, namely a total of 27%. In his subjective annotations, there were on the other hand more utterances with high or low recording volumes compared to annotator 1 (3.3% and 4.1% respectively). In continued work, we will evidently need to introduce stricter guidelines for this part of the task.

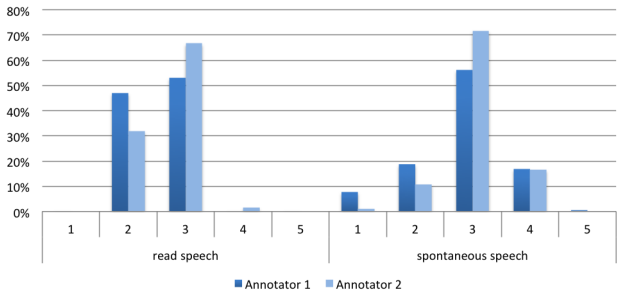

Figure 3: Fluency results read and spontaneous speech

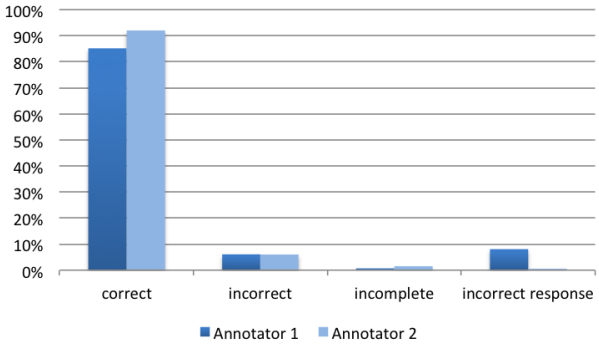

Figure 4: Grammar annotations

Most users made no or only little use of the help function, as can be seen in Figure 7. We presume that this was at least in part due to the advanced level of the test subjects, and the fact that the lessons were based on material which they had already covered in class.

Figure 8 shows the accept/reject rates individually listed for all 11 users. We can see that 7 users got a high acceptance rate, one user had about the same proportion of accepts and rejects and only 3 users had a higher rate of rejects. Looking at Table 5, it seems plausible that the poor results for these three users are largely due to issues with recording quality. For user 6, 94% of all recorded utterances contained (low or high) background noise; the same applies to user 9, with a proportion of 83% showing background noise. The problem with user 8 was mainly poor recording quality due to low volume (43%) and also some background noise (24%).

Table 6 shows an overview of the correlation between the recording quality and the acceptance rate of the system. Out of the 484 low-noise utterances for annotator 1, a proportion of 88% were accepted and 12% rejected. The proportion for high-noise was much worse, since none of these utterances were accepted by the system. The proportions

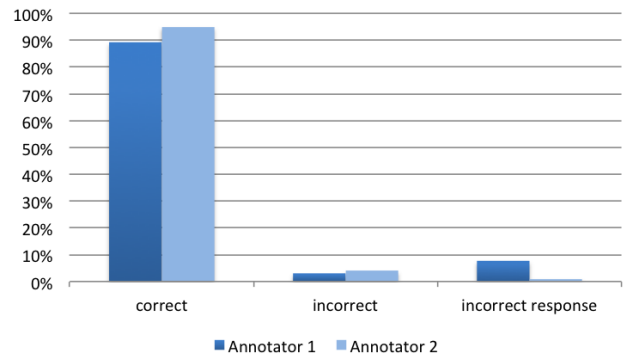

Figure 5: Vocabulary annotations

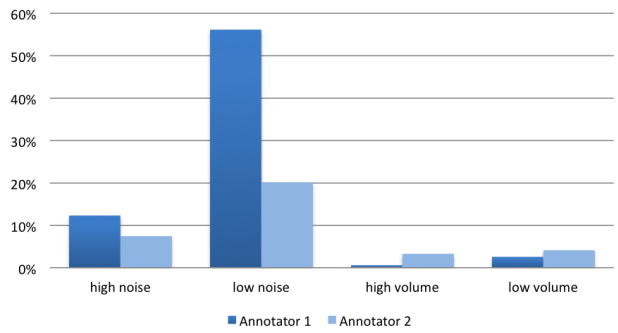

Figure 6: Recording quality annotations

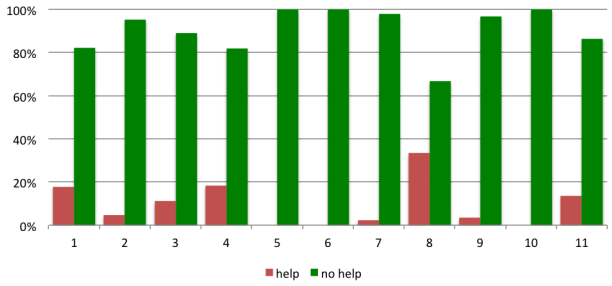

Figure 7: Usage of help function

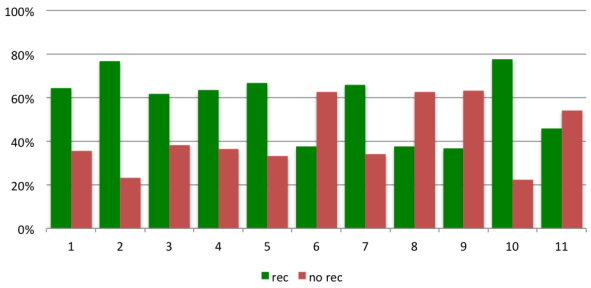

Figure 8: Accept/reject results

are similar but less extreme for annotator 2: out of 173 low-noise utterances, 70% were accepted and 30% were rejected, and again the opposite behavior applied for the 64 high-noise utterances, where only 28% were accepted and 72% rejected. Unsurprisingly, it seems clear that high background noise is an important disruptive factor which lowers the acceptance rate, whereas utterances with low background noise are most often accepted by the system. This same reasoning also applies to cut off utterances. However, since we only have very few of those in the current set of data (1%), they are not listed separately.

4. Availability of corpus data

The initial tranche of data described in the preceding section consists of 814 annotated utterances, and can be downloaded from `http://www.issco.unige.ch/en/research/resources/`. The zipfile contains the recorded wavfiles themselves, in RIFF format, and an Excel spreadsheet with the annotation data. Each line in the spreadsheet has a link to the relevant wavfile. A second tab on the spreadsheet provides demographic data – in particular, linguistic background – for the subjects used.

The intention was to run the experiment for four weeks, presenting new lessons at the rate of two a week. Based on the results of our initial experience and previous evaluation exercises of a similar nature (Rayner and Tsourakis, 2013); (Jolidon, 2013), we consider it reasonable to expect to log at least 5 000 utterances. Unfortunately, for administrative reasons beyond our control, we were forced to put data collection temporarily on hold after a few days. We expect to resume later in the Spring of 2014, and will post new annotated data as it becomes available.

We hope that this substantial annotated corpus of spontaneous child learner speech may be of interest for at least a few researchers in SLA and CALL. In the context of our own work, we have two main reasons for collecting the data. First, we wish to measure how students improve, and compare their performance against that of a parallel control class at the same school, which is not using the system. Second, we will use the data to improve the speech recognition process; in particular, we will be able to better tune the threshold parameters which control the balance between false positives and false negatives.

5. Discussion and conclusions

Although the initial dataset is on the small side, it is enough to give us an idea of how well a system like CALL-SLT can function as a corpus collection tool. There are clear strengths and weaknesses. On the positive side, corpus collection could be performed easily at a remote location. The data collected gives every appearance of being largely spontaneous. Students used the help function very little to find appropriate ways to respond, but guessed them intuitively. Most students appeared to play the game "seriously", i.e. answered so as to try to maximize their scores. We found some cases where students were more or less obviously clowning around, answering frivolously in order to amuse themselves or their classmates, but these accounted for no more than a few dozen utterances. Replaying the relevant sessions, our impression was that the students in

User	Recording quality	Proportion
6	low background noise	38%
	high background noise	56%
8	low background noise	12%
	high background noise	12%
	low volume	43%
	discard	7%
9	low background noise	37%
	high background noise	46%

Table 5: Recording quality for poorly performing users

	accept/ high	accept/ low	reject/ high	reject/ low
Annotator 1	0	425	106	59
Annotator 2	18	121	46	52

Table 6: Correlation between background noise and acceptance rate (high and low levels of background noise)

question found that frivolous responses were mostly rejected, and tired of the game[1]. This suggests to us that it is necessary for the recognition language model to be fairly strict. By way of contrast, the popular Duolingo site has extremely relaxed speech recognition feedback, usually accepting playful variants of the kind our students were using. Anecdotally, our impression is that this behaviour is correspondingly more common there.

On the negative side, the fact that we were not present while the data was being collected meant that we had no direct control over the process. An immediate consequence was high levels of background noise; a few students, who had a quiet background, appeared to be doing the exercises on their own at home, but the majority seemed to find it more enjoyable as a group activity. It was also apparent that the strictness of the language model had a downside. When responses were not accepted by the system (as far as we could see, quite often due to the noisy environment), a frequent pattern was for the student to hyperarticulate in the following utterance, making the data less natural and useful.

We are for the moment guardedly optimistic about the merits of our scheme, and hope to be able to translate the lessons learned from the initial round of experiments into improved data collection protocols when testing is resumed. We expect to report on this work in due course.

6. References

Baur, C., Rayner, M., and Tsourakis, N. (2013). A textbook-based serious game for practising spoken language. In *Proceedings of ICERI 2013*, Sevilla, Spain.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, Cambridge.

Fuchs, M., Tsourakis, N., and Rayner, M. (2012). A scalable architecture for web deployment of spoken dialogue systems. In *Proceedings of LREC 2012*, Istanbul, Turkey.

Gruenstein, A., McGraw, I., and Badr, I. (2008). The WAMI toolkit for developing, deploying, and evaluating web-accessible multimodal interfaces. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 141–148. ACM.

Jolidon, A. (2013). *Reconnaissance vocale et amélioration de la prononciation : élaboration et évaluation de leçons avec le logiciel CALL-SLT*. Masters thesis, Université de Genève.

Jurčiček, F., Keizer, S., Gašić, M., Mairesse, F., Thomson, B., Yu, K., and Young, S. (2011). Real user evaluation of spoken dialogue systems using Amazon Mechanical Turk. In *Proceedings of Interspeech 2011*, Florence, Italy.

McGraw, I., Lee, C. Y., Hetherington, I. L., Seneff, S., and Glass, J. (2010). Collecting voices from the cloud. In *Proceedings of LREC 2010*.

Morrissey, F. A., Fäs, H., Marchini, D., and Stotz, D. (2001). *Ready For English 1*. Klett und Balmer AG, Zug, Switzerland.

Rayner, M. and Tsourakis, N. (2013). Methodological issues in evaluating a spoken CALL game: Can crowdsourcing help us perform controlled experiments? In *Proceedings of the SLaTE Workshop*, Grenoble, France.

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

Rayner, M., Bouillon, P., Hockey, B.A., and Nakao, Y. (2008). Almost flat functional semantics for speech translation. In *Proceedings of COLING-2008*, Manchester, England.

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

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

Werbach, K. and Hunter, D. (2012). *For the Win: How Game Thinking Can Revolutionize Your Business*. Wharton Digital Press.

[1] An unexpected exception, found by one ingenious student, was that a Spanish *¡Hola!* was consistently accepted as an alternative to "Hello!".