7th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2017)
25-26 August 2017, Stockholm, Sweden

Overview of the 2017 Spoken CALL Shared Task

Claudia Baur[1], *Cathy Chua*[4], *Johanna Gerlach*[1]
Manny Rayner[1], *Martin Russell*[2], *Helmer Strik*[3], *Xizi Wei*[2]

[1]FTI/TIM, University of Geneva, Switzerland
[2]Department of Electronic, Electrical and Systems Engineering, University of Birmingham
[3]Centre for Language Studies (CLS), Radboud University Nijmegen
[4]Independent researcher

Claudia.Baur@unige.ch, Johanna.Gerlach@unige.ch, Emmanuel.Rayner@unige.ch,
XXW395@student.bham.ac.uk, m.j.russell@bham.ac.uk, w.strik@let.ru.nl
cathyc@pioneerbooks.com.au

Abstract

We present an overview of the shared task for spoken CALL. Groups competed on a prompt-response task using English-language data collected, through an online CALL game, from Swiss German teens in their second and third years of learning English. Each item consists of a written German prompt and an audio file containing a spoken response. The task is to accept linguistically correct responses and reject linguistically incorrect ones, with "linguistically correct" being defined by a gold standard derived from human annotations; scoring was performed using a metric defined as the ratio of the relative rejection rates on incorrect and correct responses. The task received twenty entries from nine different groups. We present the task itself, the results, a tentative analysis of what makes items challenging, a comparison between different metrics, and suggestions for a continuation.

Index Terms: CALL, shared tasks, speech recognition, metrics

1. Introduction

There are now many episodes in the history of human language technology showing that the introduction of a shared task[1] has had a positive effect on a particular area. A prominent series of examples are the various tasks based on the Wall Street Journal corpus, including speech recognition [1], parsing [2] and several types of semantic analysis [3]. Perhaps even more importantly, work on machine learning during the 21st century has to a considerable extent been driven by the handwritten digit recognition task [4]. Other well-known examples of shared tasks include ATIS in the early 90s [5], which had a strong effect on interactive spoken language systems; the Named Entity Recognition task [6], which similarly influenced work on information extraction; and the Recognizing Textual Entailment task [7], which has influenced work on question answering.

In all these cases, introduction of the shared task created a new community with frequent productive interactions between many groups, and substantially advanced a whole subfield inside the space of a few years. The sociology of the process has become familiar to many researchers. A shared task forces each group to look closely at what other groups are doing, and in particular to study methods which are achieving high scores in the competitions. It encourages development of a common vocabulary of concepts. Above all, it introduces widely accepted evaluation procedures and metrics that permit objective comparisons, both between systems developed by different groups and between different versions of single systems. It is easier to achieve progress when people agree on what "progress" consists of, and how it can be measured.

As the series of 'Speech and Language Technology in Education' (SLaTE) workshops[2] attests, speech recognition for CALL has become an established field. At the 2015 workshop in Leipzig, the authors of the present paper suggested that the time might have arrived to define a shared task for this area. The response was positive enough that we presented a paper at the LREC conference in May 2016 [8] with a concrete definition of a possible task, and released training data and resources on July 13 of the same year. Test data was released on March 13 2017, with a deadline of one week to submit results; each competing group was allowed to submit up to three entries. We were pleased to receive a total of twenty submissions from nine different groups.

In the rest of this paper, we present an overview of the shared task. §2 describes the task, data, metric and resources. §3 presents the results. In the following two sections we present some analysis: §4 considers what the data tells us about the issues that appear to make test items easier or harder, and §5 discusses the adequacy of the scoring metrics used. In conclusion, §6 makes suggestions about what to do next and §7 briefly discusses ethical issues.

2. Task, data, resources and metric

One of the most common types of spoken CALL exercise is prompt-response: the system gives the student a prompt, the student responds, and the system either accepts or rejects the response, possibly giving some extra feedback. The prompt can be of various forms, including L2 text ("read the following sentence"), L1 text ("translate the following sentence into the L2"), multimedia ("name this object"), or some kind of combination. Prompt-response exercises are for example used heavily in the popular Duolingo application.[3]

We proposed a spoken prompt-response task using data collected from an English course developed for German-speaking Swiss teenagers doing their first to third year of English [9, 10]. The course runs on CALL-SLT [11], a spoken CALL platform which has been under development at Geneva University since 2009[4]. It is based on a textbook commonly used in German-speaking Switzerland [12] and consists of eight lessons: (1) at

[1]Another common term is "competitive-collaborative task".

[2]http://hstrik.ruhosting.nl/slate/
[3]https://www.duolingo.com/
[4]http://callslt.unige.ch/demos-and-resources/

10.21437/SLaTE.2017-13

the train station, (2) getting to know someone, (3) at the tube station, (4) at the hotel, (5) shopping for clothes, (6) at the restaurant, (7) at the tourist information office, and (8) asking/giving directions. Each lesson offers an interactive dialogue permitting many variations, which allows the students to practise their oral conversational skills. The emphasis is on a communicative approach to second language acquisition, putting more weight on achieving a successful interaction than on minor grammatical or pronunciation flaws in the utterances.

Each prompt in the course is a combination of a multimedia file in the L2 (English) and a written text instruction in the L1 (German). To give a typical example, the system plays a short animated clip with an English native speaker asking the question, "How many nights would you like to stay at our hotel?" and simultaneously displays the German text, "Frag: Zimmer für 3 Nächte" (Ask: room for 3 nights). The text indicates how the student is supposed to answer in the L2. In this case, an acceptable response would be something like "I want a room for three nights", "Do you have a room for three nights?" or "I would like to stay for three nights". The intention is that a reasonably wide variety of grammatically and linguistically correct utterances are accepted, as long as they correspond to the meaning of the German prompt, so the student is able to practise spontaneous generative language skills. A response can be rejected for a variety of reasons, including incorrect use of vocabulary, grammatical incorrectness, incorrect use of the user interface, bad pronunciation, bad recognition due to insufficient recording quality, etc.

Once the student has responded, by speaking into the headset or onboard mic, the system performs speech recognition and then matches the recognised utterance against the prompt's specification of what should be counted as a correct answer. If there is a match, the system gives positive feedback by displaying a green frame around the text prompt, and moves on to the next dialogue state. If the utterance is rejected, a red frame (negative feedback) is shown and the student is asked to repeat or reformulate their response. The screenshot in figure 1 illustrates the process.

Figure 1: *CALL-SLT interface. The system has just played the English-language video (top) and shown the German-language text prompt (middle). The student asked for a help example by pressing the question-mark icon (bottom right). This is shown in the bottom pane and also has also been played in audio form. After listening to the help, the student pressed the microphone icon and spoke. The system accepted, shown by the green border, and will now move to the next dialogue state.*

The task we proposed was to simulate the ideal behavior of the above system on logged data. Results were scored against a gold standard defined by human annotators. Each item in the test-set is a pair consisting of a text prompt and a recorded audio file. The pair is to be labelled as either "accept" (the audio file represents a linguistically correct response to the text prompt), or "reject" (it does not).

2.1. Data

The core resource for the task was an English speech corpus collected with the CALL-SLT dialogue game. In total, the corpus contains 38,771 spontaneous speech acts in the form of students' interactions with the dialogue system. The data was collected in 15 school classes at 7 different schools in Germanophone Switzerland during a series of experiments in 2014 and early 2015 [10]. For the task, we made available an annotated subset of the corpus. The training corpus contained 5,222 utterances and the test corpus 996 utterances. The utterances in the training and test data sets were selected based on the following criteria with decreasing level of importance: 1) student's total number of interactions (grouped into three categories: <50, 50-200, >200 interactions), 2) pre-placement test score (out of 41), 3) gender, 4) age (ranging between 12 to 15 years). This methodology allows us to have a representative and well-balanced selection of interactions in both the training and test corpora. The two data sets contained utterances from motivated and less motivated students, from stronger and weaker students, from both male and female students and from students with different ages. Table 1 presents data selection details for both the training and test data sets.

	Training data		Test data	
		%		%
Utterances	5222		996	
Speakers	66		25	
> 200 interactions	19	28.8%	7	28.0%
50–200 interactions	35	53.0%	13	52.0%
< 50 interactions	12	18.2%	5	20.0%
Average pre-placement test score	19.18		18.44	
Female	34	51.5%	12	48.0%
Male	32	48.5%	13	52.0%
Average age	13.9		14	

Table 1: *Data selection criteria for training and test data sets*

To make the data set more interesting and challenging, short utterances such as "hello", "bye", "yes", "no" and "thanks", which occur very frequently in the corpus and are almost always well pronounced by the subjects, were removed.

The data was selected so that the sets of speakers in the test and training portions were disjoint. The test set contained data from 25 speakers, 12 female and 13 male. Table 2 presents details.

2.2. Annotated data and annotation procedure

The audio data for the training and test sets was released on the shared task site[5] together with accompanying metadata in the form of CSV-formatted spreadsheets, with one line for each

[5] http://regulus.unige.ch/spokencallsharedtask/

ID	m/f	#Utts	ID	m/f	#Utts
1	m	5	14	m	14
2	m	28	15	f	19
3	f	48	16	m	11
4	m	1	17	f	11
5	m	11	18	m	35
6	m	2	19	m	44
7	f	29	20	f	47
8	m	9	21	f	86
9	m	11	22	m	104
10	m	5	23	f	209
11	f	32	24	f	4
12	f	7	25	f	23
13	f	201			

Table 2: *Distribution of test data by speaker. "m/f" = male/female, #Utts = number of recorded utterances for speaker in test set.*

audio file. For the training data, the spreadsheet columns represented the identifier, the prompt, the name of the audio file, the transcription, and judgements (correct/incorrect) for "language" and "meaning". A judgement of "correct" in the "language" column meant that the audio file was a fully correct response to the prompt. "Incorrect" in the "language" column and "correct" in the "meaning" column meant that it was linguistically incorrect, but semantically correct (cf. Table 3). The metadata for the test set was similar, but the version released to participants omitted the transcription and judgement columns. These were kept secret until after the result submission deadline.

For the training data, we thought we would be able to carry out the annotation process efficiently by splitting it into two phases, respectively for transcription and text judging. In the first phase, audio files were transcribed by native German/Swiss German speakers fluent in English; in the second, each prompt/written response pair was judged for linguistic and semantic correctness by three native speakers of English. We hoped that by adopting this procedure and only retaining items where the judges were unanimous, we would obtain highly reliable judgements; however, careful analysis of a sample of the data convinced us that we had been too optimistic. The annotations contained more noise than we had expected, and perhaps as many as 3–4% of the judgements were incorrect.

For the test data, where this level of noise was clearly unacceptable, we rechecked the judgements after receiving the submissions. Having the submissions available made it possible for us to focus attention on the examples for which the largest number of submissions disagreed with the judgements, implying that the judgements in question were the ones most likely to be incorrect. (Details are presented in the next section). Ordering the examples in this way, by the number of submissions which gave the example the wrong label, we carefully went through the top half of the set. This meant that we checked each example where at least four submissions disagreed with the current judgement. Two of the authors, one an English native speaker and one a German native speaker fluent in English, carefully listened to the examples together and discussed each one until they had reached clear agreement. The principles used to adjudicate borderline examples (in all cases according to the principle "other things being equal") were as follows:

- Examples with background noise, crosstalk or interruptions were counted as linguistically and semantically correct.
- Disfluencies and repetitions were counted as linguistically and semantically correct.
- Mispronunciations were counted as linguistically and semantically correct, unless the mispronunciation either a) resulted in a different English word meaningful in the given context or b) was incomprehensible.
- Foreign words (usually German/Swiss German but occasionally French) were counted as linguistically incorrect, but semantically correct if they were close enough to the corresponding English word.

2.3. Resources

In order to make it easier for groups to attempt the proposed task, we provided some other resources: a baseline Kaldi recogniser for the domain, a baseline domain grammar, sets of speech data preprocessed though two domain recognisers, and a baseline Python script that combined the other resources to perform the task. All these materials were made available for free download from the shared task site. We present brief descriptions of these resources below.

2.3.1. Baseline Kaldi recogniser

For people who wanted to experiment with recognition methods, we included acoustic models, language models and scripts for Kaldi [13], a state-of-the-art open-source recogniser platform. Three baseline deep neural network (DNN)-hidden Markov model (HMM) hybrid systems were trained on the training part of WSJCAM0 [14] (WSJCAM0_TR) and 90% of the training corpus of the Spoken CALL Shared Task (90ST). Base_DNN1, Base_DNN2 and Base_DNN3 were trained on 1. WSJCAM0_TR, 2. WSJCAM0_TR plus 90ST, and 3. 90ST, respectively. A Gaussian mixture model (GMM)-HMM system is initially trained to provide the triphone tree and the state level time alignment. The hidden layers of the DNN are unsupervised pre-trained as a stack of restricted Boltzmann machines (RBMs). The output layer is initialized randomly. The network has 4 hidden layers with 1024 neurons in each layer and a softmax output layer with 1516 units. A basic bigram language model, trained on the Shared Task training data, was also included. Recognition experiments on the remaining 10% of the Shared Task data (Table 4), show that Base_DNN1 and Base_DNN2 with WSJCAM0 performed worse than Base_DNN3. This may be because the majority of the training set was from WSJCAM0, and the adult speech from WSJCAM0 is not like the speech from the task. To make the system more like the shared task testing data, we further trained the network from the Base_DNN2 system with 90ST to obtain DNN4. With 13.92% word error rate, DNN4 outperforms the other three baseline systems. After ten cross-validation experiments, this method of adaption was chosen to train our final baseline system for the shared task.

2.3.2. "Pre-recognised" versions of data

For the benefit of groups who only wished to explore the language processing aspects of the task, we processed test and training data through both the baseline Kaldi recogniser and a Nuance Toolkit recogniser with a language model created from the baseline response grammar, and supplied versions of the task metadata which included the recognition results produced.

Prompt	Transcription	language	meaning
Frag: Zimmer für 6 Nächte	I would like a room for six nights	correct	correct
Frag: Zimmer für 6 Nächte	I wants a room for six nights	incorrect	correct
Frag: Zimmer für 6 Nächte	I want a room for five nights	incorrect	incorrect
Frag: Zimmer für 6 Nächte	It's raining outside	incorrect	incorrect

Table 3: *Examples showing use of the "language" and "meaning" judgements in the annotated data. The prompt, "Frag: Zimmer für 6 Nächte" means "Request: room for 6 nights".*

system	training data	WER
Base_DNN1	wsjcam0	72.12
Base_DNN2	wsjcam0+90ST	19.62
Base_DNN3	90ST	16.61
DNN4	90ST	13.92

Table 4: *Results(% WER) of DNN baseline systems*

2.3.3. Baseline response grammar

We also provided a version of the existing CALL-SLT response grammar, which contained 565 prompts with a total of 43,862 possible responses. The grammar was supplied in a minimal XML format, where each item consisted of the original German text prompt, an English translation of the prompt, and a list of possible responses. A typical record from the grammar is shown in Figure 2. We stated clearly that the response grammar **was not intended to be exhaustive**. The task is open-ended; ideally, the system should accept any grammatically correct, adequately pronounced response which corresponds to the prompt, and the grammar only gives plausible examples of such responses. Since the grammar was automatically derived from the one used to perform the actual data collection, we knew that it gave useful coverage, but it could evidently be improved.

2.3.4. Baseline system

Finally, we supplied a simple Python script which instantiated a minimal example of a system capable of performing the shared task. The script reads a "pre-recognised" set of metadata (§2.3.2) and the baseline response grammar (§2.3.3) and produces a CSV-formatted spreadsheet of accept/reject decisions, labelling a response as "accept" if and only if the recognition result is in the response grammar.

2.4. Metrics

For reasons explored in our 2016 paper, we are not convinced that standard metrics such as the F-measure are appropriate for scoring tasks like the one we proposed here. In particular, one serious objection is that these metrics can give a higher score to a system which always accepts than they do to a system which behaves in a normal way, accepting correct responses more frequently than incorrect ones. Intuitively, a system which always accepts is useless; if a system is useful, it is precisely *because* it responds differently to correct and incorrect responses. We consequently decided to base our metric directly on the idea of measuring the degree of difference between the system's behaviour in the case of the two types of response. A second intuition was that false accepts are not all equally serious. A false accept of a response which is linguistically incorrect, but has the right meaning, should be less serious than a false accept of a response which is not even semantically correct. We call these two kinds of false accepts respectively "plain false accepts" and "gross false accepts".

Slightly adapting the treatment of [15], we define metrics as follows. We assume that we are given a set of annotated prompt/response interactions, where in each case the annotations show whether the response was correct or incorrect, both linguistically and semantically, and whether it was accepted or rejected. We write CA for the number of of correct accepts, CR for the number of correct rejects, FA_1 for the number of plain false accepts, FA_2 for the number of gross false accepts and FR for the number of false rejects. We set

$$FA = FA_1 + k.FA_2$$

for some constant k, weighting gross false accepts k times more heavily than plain false accepts, and

$$Z = CA + CR + FA + FR$$

Then we write $C_A = \frac{CA}{Z}$, $C_R = \frac{CR}{Z}$, $F_A = \frac{FA}{Z}$, $F_R = \frac{FR}{Z}$ and define metrics in terms of the four quantities C_A, C_R, F_A, F_R, which total to unity. Looking first at traditional metrics, we consider precision ($P = \frac{C_A}{C_A+F_A}$), recall ($R = \frac{C_A}{C_A+F_R}$), F-measure $F = \frac{2PR}{P+R}$) and scoring accuracy ($SA = C_A + C_R$). Scoring accuracy SA is related to classification error E by the equation $SA = 1 - E$, and maximising SA is equivalent to minimising E.

Generally, all of the above metrics are based on the idea of minimising some kind of error. In contrast, D, the metric based on differential response which we used for the task, is defined as the ratio of the relative correct reject rate (the reject rate on incorrect responses) to the relative false reject rate (the reject rate on correct responses). We put $RC_R = \frac{C_R}{C_R+F_A}$ and $RF_R = \frac{F_R}{F_R+C_A}$, then define

$$D = \frac{RC_R}{RF_R} = \frac{C_R/(C_R+F_A)}{F_R/(F_R+C_A)} = \frac{C_R(F_R+C_A)}{F_R(C_R+F_A)}$$

We announced when proposing the task that results would be ranked in terms of D with $k = 3$, but we also present scores for the other metrics.

3. Results

We received twenty entries from nine different groups. Of these, ten entries (four groups) used the pre-recognised baseline Kaldi data; four entries (three groups) used the pre-recognised baseline Nuance data; and six entries (three groups) used their own recognisers. One group submitted entries both for their own recogniser and for Nuance data. Table 5 present submissions and scores using D and several other metrics, also comparing with three baseline systems built using the resources described in §2.3.

```
<prompt_unit>
   <prompt>Frag : Wie viel kostet es ?</prompt>
   <translatedprompt>Ask: How much does it cost?</translatedprompt>
   <response>how much does it cost</response>
   <response>how much does this cost</response>
   <response>how much is it</response>
   <response>how much is this</response>
</prompt_unit>
```

Figure 2: *XML reference grammar example.*

Id	Rec	Pr	R	F	SA	RCR	RFR	D
KKK	Custom	0.881	0.845	0.862	0.840	0.739	0.155	**4.766**
JJJ	Custom	0.871	0.848	0.859	0.838	0.717	0.152	4.710
BaselinePerfectRec	n/a	0.995	0.781	0.875	0.839	0.989	0.219	4.512
CCC	Kaldi	0.687	0.933	0.791	0.801	0.300	0.067	4.468
III	Custom	0.752	0.904	0.821	0.805	0.421	0.096	4.371
MMM	Kaldi	0.732	0.905	0.809	0.818	0.413	0.095	4.353
OOO	Kaldi	0.739	0.899	0.812	0.818	0.430	0.101	4.273
GGG	Kaldi	0.639	0.957	0.766	0.769	0.173	0.043	3.998
AAA	Kaldi	0.606	0.973	0.747	0.749	0.098	0.027	3.678
FFF	Kaldi	0.631	0.951	0.759	0.762	0.164	0.049	3.352
BBB	Kaldi	0.748	0.862	0.801	0.798	0.461	0.138	3.335
HHH	Kaldi	0.602	0.971	0.743	0.747	0.096	0.029	3.289
PPP	Custom	0.838	0.795	0.816	0.786	0.660	0.205	3.217
NNN	Kaldi	0.713	0.880	0.788	0.790	0.383	0.120	3.188
QQQ	Custom	0.854	0.753	0.800	0.770	0.713	0.247	2.882
DDD	Kaldi	0.777	0.811	0.794	0.779	0.539	0.189	2.857
LLL	Nuance	0.814	0.754	0.783	0.746	0.623	0.246	2.533
BaselineNuance	Nuance	0.822	0.723	0.770	0.731	0.652	0.277	2.358
EEE	Nuance	0.816	0.729	0.770	0.731	0.636	0.271	2.347
SSS	Nuance	0.737	0.820	0.776	0.743	0.423	0.180	2.346
RRR	Custom	0.884	0.584	0.703	0.668	0.818	0.416	1.965
BaselineKaldi	Kaldi	0.957	0.439	0.602	0.588	0.951	0.561	1.694
TTT	Nuance	0.622	0.897	0.735	0.713	0.148	0.103	1.437

Table 5: *Results for 20 anonymised submissions and three baseline systems. "Rec" = recogniser used ("Kaldi" = pre-recognised Kaldi, "Nuance" = pre-recognised Nuance, "Custom" = own recogniser), "Pr" = precision, "R" = recall "F" = F-measure, "SA" = scoring accuracy, "RCR" = relative correct rejections, "RFR" = relative false rejections, "D" = D-measure. "Baseline Kaldi" = system with baseline Kaldi recogniser and baseline XML grammar; "Baseline Nuance" = system with baseline Nuance recogniser and baseline XML grammar; "Baseline PerfectRec" = system with input from transcriptions and baseline XML grammar.*

4. What makes items difficult?

In order to study the degree of difficulty of the test items, we combined the results of all 20 submitted entries, with those of the "Nuance" and "Kaldi" baseline systems, making a total of 22 entries. A full table is available on the "Test data" tab of https://regulus.unige.ch/spokencallsharedtask/. Table 6 presents a summary and some representative examples. The greater part of the test set, as can be seen, was quite straightforward; 581 of the 996 items end up in the "Easy" group, and were scored correctly by at least 17 of the 22 entries. (232 were scored correctly by all entries). Another 265 items, the "Intermediate" group, were scored correctly by more than half of the entries. The remaining 150 items were clearly challenging, with 24 being scored correctly by only five or fewer entries.

In order to gain some understanding of a few obvious factors which might make items easier or harder, we performed an annotation of the test data. Three of the authors (a native Swiss German speaker fluent in English, a native English speaker with some German, and a native English speaker with no German) listened to each audio file separately using an online tool and categorized them on the following five scales:

Crosstalk Could you hear anyone other than the student talking? (yes/no)

Non-speech noise Could you hear any non-speech noises, for example background noise, breath noise, etc? (no/weak/strong; "weak" was defined as "clearly softer than the speech" and "strong" as "comparable in loudness with the speech").

Stuttering/repetition Did the student stutter, repeat themself, or in some other way clearly change their mind about what they were going to say? (yes/no)

Incomprehensible Was any word spoken by the student incomprehensible to you? (yes/no)

Faint Was the volume of the student's speech clearly much fainter than usual? (yes/no)

Id	Prompt	Transcription	language	meaning	#Bad
Difficult (11–22 entries wrong, 150 items)					
3709	Sag: Ich habe 2 jüngere Brüder	i have two young brothers	incorrect	incorrect	22
3944	Frag: Hosen	i want pants	correct	correct	20
4271	Frag: ein T-Shirt	i will like a t-shirt	incorrect	correct	19
4519	Frag: ein Ticket zum Green Park	can i have a ticket to the green park	incorrect	correct	15
4080	Frag: mein Steak rare	i want a rare steak	correct	correct	13
4683	Frag: Gibt es einen Lift?	is there a ascenseur	incorrect	incorrect	11
Intermediate (6–10 entries wrong, 265 items)					
4540	Sag: Kann ich mit Kreditkarte bezahlen	i like to pay with credit card	incorrect	correct	9
4155	Frag: Wo kann ich ein Shampoo kaufen?	where i can buy shampoo	incorrect	correct	6
3877	Frag: Erbsen	i would like some peas	correct	correct	6
Easy (0–5 entries wrong, 581 items)					
4679	Frag: Doppelzimmer	can i have a double room	correct	correct	4
4419	Frag: ein Ticket für Mamma Mia	i want one ticket for mamma mia	correct	correct	2
4085	Frag: Ich möchte die Dessertkarte	i would like the dessert card	incorrect	correct	0

Table 6: *Examples of prompt/response pairs with different levels of difficulty, estimated by the number of entries assigning the wrong label. "Id" = identifier of utterance in Shared Task test set (clickable link to audio file), "Prompt" = German language prompt, "Transcription" = transcription of student response, "language" = judgement of responses's correctness in terms of both language and meaning, "meaning" = judgement of response's correctness only in terms of meaning, "#Bad" = number of submissions out of 22 making incorrect decision.*

The categories were chosen as labeling conditions which a) occurred reasonably often in the data, much of which was recorded in noisy environments, b) could reasonably be expected to make the accept/reject decision difficult and c) were easy to judge. Unfortunately, we were not able to include any measure of pronunciation quality; previous experience had convinced us that we would not be able to judge this usefully in the time available. We found very poor inter-annotator agreement on "Non-speech noise", where about half the data was annotated as "Weak" by each judge. We consequently collapsed "No" and "Weak" together for this scale, making all five scales binary. The judgements from the three annotators were combined using majority voting; Table 7 shows agreement between annotators. We also added a sixth category, out-of-vocabulary (OOV), which was computed automatically; an example in the test data was counted as OOV if at least one word in the transcription failed to occur in either the training data or the baseline response grammar (cf. §2.3.3). Finally, we computed the number of word errors (sum of insertions, deletions and substitutions) for each item, using the recognisers for which we had available data. These were the baseline Kaldi and Nuance recognisers, and the recogniser for the JJJ entry, whose author submitted them to us to be made publicly available on the shared task site. Table 8 shows the distribution of the resulting metrics over the different bands from Table 6.

We draw the following tentative conclusion from this data. We see that for all 6 measures, the percentages are lower for the easy cases (0–5 labelling errors). The relative differences are smaller for Faint, and larger for OOV and the other four properties of the utterances (CT, NSN, Stut, Inc). Thus it seems that esp. the latter 5 measures have a substantial effect on item difficulty. At the same time, the fact that only 38 percent of the "Difficult" items are marked for any of the categories suggests that other factors might be relevant. Obvious candidates are of course pronunciation related measures, at segmental and/or prosodic level. Second, looking at the word error rate columns, we see that WER for all three recognisers is much lower for the "Easy" group; this supports the commonsense hypothesis that recognition quality is an important factor in determining whether an item is easy or difficult. Conversely, the fact that the WER is only slightly higher in the "Difficult" group than it is in the "Medium" group also suggests that other factors might be involved.

These intuitive impressions are strengthened by carrying out a basic ANOVA analysis. The effect of word error on labelling error is very significant ($p < 10^{-15}$) for the entries using recognisers where WER data is available, but WER still accounts for less than a quarter of the variance in the labelling error for any recogniser. ANOVA also shows strongly significant effect on labelling error from the categories OOV, Crosstalk and Stuttering/Repetition ($p < 10^{-8}$, $p < 10^{-4}$ and $p < 10^{-3}$ respectively), but these categories account for only a few percent of the variance in the labelling error. This supports the intuitive conclusion that these three categories, while significant, are not central to the task.

Category	Agree	κ
Crosstalk	0.974	0.767
Non-speech noises	0.548	0.233
Non-speech noises (collapsed)	0.895	0.275
Stuttering/repetition	0.971	0.635
Incomprehensible	0.885	0.075
Faint	0.775	0.377

Table 7: *Results of annotation on test set by three judges: proportion of data agreeing and Light's κ scores.*

5. Metrics

We calculated the correlation between metrics using Kendall's τ, a common statistic for measuring similarity of ordinal sequences. Two metrics have a Kendall's τ of 1 on a set if they put the elements of the set in the same order, -1 if they put them in reverse order. Table 9 summarises the results.

#Bad	CT	NSN	Stut	Inc	Faint	OOV	$\geqslant$1	Word errors		
								Nuance	Kaldi	JJJ
0–5	1.4	1.2	1.0	1.5	12.2	3.1	16.7	0.938	0.676	0.411
6–10	4.9	3.4	5.3	1.1	16.6	11.7	34.7	2.626	2.211	1.242
11–22	7.3	2.0	4.7	2.7	14.7	16.7	38.0	2.833	2.273	1.580
All	3.2	1.9	2.7	1.6	13.8	7.4	24.7	1.673	1.325	0.808

Table 8: *Possible indicators of difficulty, broken down by number of entries out of 22 assigning the wrong label. "#Bad" = number of entries assigning wrong label. Left-hand side: percentage of items presenting six types of possible problems. "CT" = crosstalk, "NSN" = strong non-speech noise, "Stut" = stuttering, repetition etc, "Inc" = at least one incomprehensible word, "Faint" = low volume in speech, "OOV" = at least one out of vocabulary word, "$\geqslant$1" = at least one of preceding factors. Right-hand side: average number of recogniser word errors per utterance for each group, for the three recognisers where data was available.*

First of all, Precision and Recall correlate negatively with each other; which isn't surprising, as there is in general a trade-off between these two metrics. Below we will focus on the three remaining metrics, the "main metrics": D, SA, and F.

We find that the D-metric used for the shared task correlates well with the scoring accuracy (hence also with the simple error rate, since the scoring accuracy is 1 minus the error rate), but rather less well with the F-measure. Interestingly, of the three main metrics, for SA the highest correlations with the other two are observed, while the correlation of D and F are lower.

Looking at Table 5, it is comforting to see that of all submitted twenty entries, the same entry, KKK, was best according to all three of the main metrics. We are however struck by the fact that the baseline "perfect-recognition" system slightly outperformed KKK on F-score, but not on D-score. We are not certain what interpretation to put on these results, and welcome discussion.

	D	SA	F	Rec	Pr
D	*	0.731	0.557	0.360	–0.036
SA	0.731	*	0.763	0.217	0.091
F	0.557	0.763	*	–0.004	0.328
Rec	0.360	0.217	–0.004	*	–0.676
Pr	–0.036	0.091	0.328	–0.676	*

Table 9: *Correlations between different metrics calculated using Kendall's τ. "D" = D-metric, "SA" = scoring accuracy (inverse of error rate), "F" = F-measure, "Rec" = recall, "Pr" = precision.*

6. Conclusions and further directions

The first spoken CALL shared task has been quite successful: we were pleased to see such an enthusiastic response from the community. We outline some ideas about where to go next.

6.1. Relative importance of speech and language processing

The results and analysis presented in §3 suggest, as expected, that speech and language processing are both essential parts of the task. One piece of evidence is that recogniser word error only accounts for a modest proportion of the variance in the labelling error. Another is that groups focusing on language processing were able to achieve results nearly as good as those who focused on speech processing; the CCC entry, which used the baseline Kaldi recogniser, obtained scores only marginally lower than those of the JJJ entry, despite the fact that the JJJ recogniser's WER was much better.

This raises the interesting prospect of creating hybrid systems which combine one group's speech processing with another group's language processing; it seems plausible to hope that a combination of this kind could produce results better than either group is obtaining at the moment. As previously noted, the ASR results for the JJJ entry have already been made publicly available.

6.2. Searching for missing factors

The quality of pronunciation (including suprasegmental/prosodic aspects) is intuitively a very important factor when determining the extent to which an item is challenging for a CALL system to handle. The results from §3 are compatible with this hypothesis; the factors we examined only account for a small proportion of the variance in the labelling error, so some large influence must be missing. It would however be very good to have some kind of *direct* evidence that the missing factors are related to pronunciation, and attempt to quantify their contribution. Groups with experience in phonetic transcription may find this an interesting topic for further investigation.

6.3. A continuation of the shared task

Finally, it is natural to think about possible continuations of the shared task itself. Here we consider four options: a) nothing, this was a one-off; b) the same task again, with new test data; c) use the same data, but define a more challenging task; d) a completely new task. Our suggestion is that we start discussing this at the SLaTE 2017 workshop, and e.g. make democratic decisions about it.

6.3.1. No continuation

Given the substantial number of entries for the task, it seems logical to follow it up. Of course, this entirely depends on continued interest. We will try to make an inventory of parties interested in a continuation, starting by asking people present at the SLaTE 2017 workshop.

6.3.2. Same task, new test data

The simplest alternative is to repeat the current task using new test data. We have large amounts of task data logged (cf. §2.1), and it would be easy to annotate a thousand more utterances.

Our feeling is that the second edition of the task would probably be best scheduled as a special session at a 2018 conference. The most plausible conferences in 2018 seem to us to

be the following:

- Interspeech 2018 (Hyderabad, India)
- LREC 2018 (Miyazaki, Japan)
- Building Educational Applications 2018 (location to be announced)

Another option might be a SLaTE workshop in 2019, which could be a satellite of Interspeech 2019 in Graz.

6.3.3. More challenging version of current task

Ideally, it would be desirable to redefine the task in some way so that pronunciation quality could be taken into account. The practical problem is that this would appear to require phonetic annotation of both the training and the test data, i.e. a minimum of 6,000 utterances.

The Geneva group, who have so far taken responsibility for data annotation, do not have resources to do this, so another partner would need to get involved.

6.3.4. New task

The current task is perhaps a little too easy (in particular, its vocabulary is only about 450 words), and there is an argument for moving to a more challenging one. This would require relevant data, including the required annotations. If you have ideas and/or data for such a new task, please contact us.

7. Ethical considerations

The ethics of shared tasks has recently been the subject of some attention. We briefly address the issues raised by Escartin and her colleagues [16], considering conflicts of interest, anonymity, gaming of the system and the balance between competitiveness and collaboration.

To start with the more obvious items, there was a potential conflict of interest in that one of the competing groups (Birmingham) was also involved in organising the task. To avoid any possibility of giving the Birmingham group an unfair advantage, we processed all the data at the Geneva site, only making it available to Birmingham according to the normal schedule. With regard to anonymity, all the results have been presented under anonymised IDs. Each group has been given the key to their own results, so they are free to choose whether to stay anonymous or reveal their identity. We were concerned at one point that the D metric (§5) could be "gamed" by a strategy which only rejected items that were almost certainly wrong. This would give a low correct reject rate and an even lower false reject rate, producing a high D. If any submission tried to use this strategy, it did not succeed. It is however a potential weakness of the D metric that we should bear in mind if the task is repeated.

More generally, as Escartin *et al* point out, the primary ethical challenge in a shared task is to encourage openness and sharing of results and discourage inappropriate competitiveness. We have been gratified to see that all participants are in fact displaying a constructive and positive attitude. It has been a pleasure to organise this event.

8. Acknowledgements

Work on the shared task at the University of Geneva was partially funded by the Swiss National Science Foundation (SNSF) under project 105219_153278. We would like to thank Karën Fort for helpful comments on ethical issues.

9. References

[1] L. R. Bahl, S. Balakrishnan-Aiyer, J. Bellgarda, M. Franz, P. Gopalakrishnan, D. Nahamoo, M. Novak, M. Padmanabhan, M. A. Picheny, and S. Roukos, "Performance of the IBM large vocabulary continuous speech recognition system on the ARPA Wall Street Journal task," in *Proceedings of ICASSP 1995*. IEEE, 1995, pp. 41–44.

[2] S. Riezler, T. King, R. Kaplan, R. Crouch, J. Maxwell, and M. Johnson, "Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (demo track)*, Philadelphia, PA, 2002.

[3] S. S. Pradhan, E. Loper, D. Dligach, and M. Palmer, "Semeval-2007 task 17: English lexical sample, SRL and all words," in *Proceedings of the 4th International Workshop on Semantic Evaluations*. Association for Computational Linguistics, 2007, pp. 87–92.

[4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016.

[5] V. Zue, S. Seneff, J. Polifroni, M. Phillips, C. Pao, D. Goddeau, J. Glass, and E. Brill, "Pegasus: A spoken language interface for on-line air travel planning," in *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, 1994, pp. 201–206.

[6] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. Association for Computational Linguistics, 2003, pp. 142–147.

[7] I. Dagan, O. Glickman, and B. Magnini, "The PASCAL recognising textual entailment challenge," in *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*. Springer, 2006, pp. 177–190.

[8] C. Baur, J. Gerlach, M. Rayner, M. Russell, and H. Strik, "A shared task for spoken CALL?" in *Proceedings of LREC 2016*, Portorož, Slovenia, 2016.

[9] C. Baur, M. Rayner, and N. Tsourakis, "A textbook-based serious game for practising spoken language," in *Proceedings of ICERI 2013*, Seville, Spain, 2013.

[10] C. Baur, "The potential of interactive speech-enabled CALL in the Swiss education system: A large-scale experiment on the basis of English CALL-SLT," Ph.D. dissertation, University of Geneva, 2015.

[11] 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.

[12] F. A. Morrissey, H. Fäs, D. Marchini, and D. Stotz, *Ready for English 1*. Zug, Switzerland: Klett und Balmer AG, 2006.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[14] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJ-CAM0: A British English corpus for large vocabulary continuous speech recognition," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 1994.

[15] S. Kanters, C. Cucchiarini, and H. Strik, "The goodness of pronunciation algorithm: a detailed performance study," *SLaTE*, vol. 2009, pp. 2–5, 2009.

[16] C. P. Escartın, W. Reijers, T. Lynn, J. Moorkens, A. Way, and C.-H. Liu, "Ethical considerations in NLP shared tasks," in *Proc. First ACL Workshop on Ethics in Natural Language Processing*, Valencia, Spain, 2017.