SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education
20-21 September 2019, Graz, Austria

Overview of the 2019 Spoken CALL Shared Task

Claudia Baur[1], *Andrew Caines*[2], *Cathy Chua*[3], *Johanna Gerlach*[1],
Mengjie Qian[4], *Manny Rayner*[1], *Martin Russell*[4], *Helmer Strik*[5], *Xizi Wei*[4]

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

claudia.baur@bluewin.ch, apc38@cam.ac.uk, cathyc@pioneerbooks.com.au,
Johanna.Gerlach@unige.ch, MXQ486@student.bham.ac.uk, Emmanuel.Rayner@unige.ch,
m.j.russell@bham.ac.uk, w.strik@let.ru.nl, XXW395@student.bham.ac.uk

Abstract

We present an overview of the third edition of the Spoken CALL Shared Task. 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" defined by a gold standard derived from human annotations. Scoring was performed using a metric based on the idea of maximising the ratios correct-accept-rate/false-accept-rate and correct-reject-rate/false-reject-rate. The third edition received sixteen entries, with the best score substantially improving on last year's edition of the task. We analyse factors which make it difficult to label items correctly, concluding that, as in the previous edition, good speech recognition is most important. Finally, we suggest a strategy for continuing the task.

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

1. Introduction

The Spoken CALL Shared Task is a series of open challenges jointly organised by the University of Geneva, the University of Birmingham, Radboud University and the University of Cambridge[1]. The task is based on data collected from a speech-enabled online tool which has been used to help young Swiss German teens practise skills in English conversation. Items are prompt-response pairs, where the prompt is a piece of German text and the response is a recorded learner English audio file. The task is to label pairs as "accept" or "reject", accepting responses which are grammatically and linguistically correct to match a set of hidden gold standard answers as closely as possible. Results are scored using metrics which reward maximisation of the difference between the system's reaction to correct and incorrect student responses.

Training data for the first edition of the Spoken CALL Shared Task was released in July 2016 and test data in March 2017 [1], attracting 20 entries. Results and seven papers were presented at the SLaTE workshop in August 2017 (http://regulus.unige.ch/spokencallsharedtask; [2, 3, 4, 5, 6, 7, 8]). Results for the second edition of the task were presented at an Interspeech 2018 special session, with 18 entries and six papers (https://regulus.unige.ch/spokencallsharedtask_2ndedition; [9, 10, 11, 12, 13, 14]).

In both the first and the second editions, scoring was performed using the D metric, formally defined in §2.3. D focuses attention on "reject" responses by measuring the ratio between the frequencies of correct rejects and false rejects. We had two reasons for doing this. First, students are much more concerned about false rejects than false accepts; this is abundantly clear in our data, when subjects can often be heard mocking or insulting the system when it incorrectly rejects them. Second, the task of optimising D is reasonably approachable, since the key task is to lower the false reject rate on correct responses. For reasons described in detail in [9], and also in §4 here, correct responses are much easier to process than incorrect responses.

By the second edition of the task, the best entries had found strong strategies for optimising D, and this metric had clearly topped out. For the third edition, the subject of the present paper. we consequently decided to move to a new and more challenging metric, D_{full}, which places equal weight on processing of correct and incorrect responses.

The rest of the paper is structured as follows. §2 describes the data, resources and metrics, in particular presenting formal definitions of D and D_{full}. §3 presents the results of the 2019 edition of the Shared Task. §4 discusses the factors which make it hard to label responses correctly. The final section concludes and suggests further directions.

2. Data, resources and metric

2.1. Data

The training data provided for the third edition of the task was the same as for the second, and is available for download from the task site (https://regulus.unige.ch/spokencallsharedtask_3rdedition). It is fully described in [2, 9], but for completeness we briefly summarise.

All data was logged during 2014 and 2015 using an online English course app developed for German-speaking Swiss teenagers in their second or third year of learning English [15, 16, 17, 2]; the app consisted of eight interactive lessons on themes like "booking a hotel room", "ordering a meal at a restaurant" and similar. The content was closely based on a textbook commonly used in germanophone Switzerland [18] and was developed in collaboration with a teacher at one of

[1] Work at Cambridge was funded by Cambridge Assessment. Work at Geneva was partially funded by the Swiss National Science Foundation under grant IZCOZ0_177065.

10.21437/SLaTE.2019-1

the schools concerned. The format was prompt/response, with prompts consisting of a combination of an English video prompt (e.g. a cartoon desk clerk asking “How can I help you?”) and a piece of text in telegraphic German describing the correct response (*Frag: Zimmer für 3 Nächt* = “Request: room for three nights”). The intention is that the student can respond freely with phrases like “I want a room for three nights”, “room for three nights please”, etc); online spoken help can be accessed if desired. Prompts were collected together into simple dialogues, with a non-deterministic progression. The app was used by about 220 students, aged between 12 and 15 years, attending 15 different classes at seven schools, and logged a total of 38,771 spontaneous speech acts.

Training data for the first edition of the task consisted of 5,222 transcribed and annotated utterances; each utterance was stored together with its associated prompt and annotations specifying a) whether it was a fully correct response to the prompt in terms of vocabulary, grammar and meaning b) whether it was correct in terms of meaning only. Annotation was performed by three human judges. For the second edition, a further 6,998 annotated utterances were created in the same format. The availability of several high-quality entries for the first edition meant that the annotation procedure could be partially automated, which also improved the quality. Test data for the first edition of the task consisted of 996 utterances, and for the second edition of 1000 utterances. In both cases, we used speakers not appearing in previously released training or test data.

For the third edition of the task, we only released new test data. We began by selecting 1785 utterances from 25 speakers who once again had not appeared in previously released data. The selected data was transcribed and separately annotated by three native speakers of English using an online tool. For each item, the annotator could read both the original German prompt and an English translation, and listen to the response. They were asked to answer the following binary questions, using “radio-button” style menus:

Fully correct Was the recorded speech a fully correct response to the prompt, in terms of grammar, vocabulary and meaning? The judges were instructed to ignore pronunciation, except if words were mispronounced to an extent that made them incomprehensible or comprehensible as a different word in the dialogue context.

Semantically correct Was the recorded speech a fully correct response to the prompt in terms of meaning, though not necessarily in terms of grammar and vocabulary?

Incomprehensible Was any word spoken by the student incomprehensible to you?

Pronunciation Was any word spoken by the student clearly mispronounced, in the sense that at least one English sound was clearly substituted by a different and incorrect English sound?

Stuttering/repetition Did the student stutter, repeat themself, or in some other way clearly change their mind about what they were going to say?

Crosstalk Could you hear anyone other than the student talking?

Strong non-speech noise Could you hear any non-speech noises, for example background noise, comparable in loudness with the student’s speech?

Faint Was the volume of the student’s speech clearly much fainter than usual?

Information for the first two categories was used in the task itself, and the remaining six in the experiments described in §4.

Items where the three judges were not unanimous on the first two categories were discarded. Incorrect responses were under-represented in the utterances that were left, so we kept all 260 “incorrect” responses, then randomly selected 740 “correct” responses to produce a test set of 1000 items. This had almost the same balance between correct and incorrect as the second edition test set, 740–260 as compared with 750–250.

2.2. Other resources

As in the second edition, the University of Birmingham group made available the Kaldi recogniser and the response grammar used for their winning entry in the first edition of the task, to act as the baseline. Both resources are described in detail in [3]. For the benefit of groups who only wished to explore the language processing aspects of the task, we processed test and training data through the baseline recogniser, and supplied versions of the task metadata which included the recognition results produced. Finally, we supplied a Python script which instantiated a minimal example of a system capable of performing the shared task, using the “pre-recognised” set of metadata and the baseline response grammar.

2.3. Metrics

We define D, D_{full} and other metrics as follows. There are three main intuitions. First, D_{full} should measure the difference between the system’s reaction to correct and incorrect responses; second, it should handle the two kinds of responses symmetrically; third, it should give a larger penalty to a false accept if the response is semantically as well as syntactically incorrect. We call false accepts of semantically incorrect responses “gross false accepts” and false accepts of semantically correct responses “plain false accepts”.

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 syntactically and semantically, and whether it was accepted or rejected. We write CA for the number of correct accepts, CR for the number of correct rejects, PFA for the number of plain false accepts, GFA for the number of gross false accepts and FR for the number of false rejects. We set $FA = PFA + k.GFA$ 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 ($SA = C_A + C_R$).

Generally, all of the above metrics are based on the idea of minimising some kind of error. In contrast, the metrics used for the Spoken CALL shared task are based on *maximising* the difference between the system’s reaction to correct and incorrect responses. We start by defining D, the metric we used for the first and second editions of the task, to be 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

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

Id	Rec	Pr	R	F	SA	RCR	RFR	D	D_A	D_{full}
BaselinePerfectRec	Text	0.977	0.907	0.940	0.916	0.940	0.093	10.080	15.075	12.327
GGG	Speech	0.901	0.935	**0.918**	**0.879**	0.736	0.065	11.348	**3.544**	**6.342**
HHH	Speech	0.884	0.946	0.914	0.873	0.689	0.054	**12.750**	3.043	6.229
III	Speech	0.883	0.945	0.913	0.871	0.688	0.055	12.416	3.027	6.130
(OOO)	Speech	0.895	0.923	0.909	0.867	0.724	0.077	9.401	3.346	5.608
(PPP)	Speech	0.895	0.923	0.909	0.867	0.724	0.077	9.401	3.346	5.608
(NNN)	Speech	0.896	0.919	0.907	0.865	0.726	0.081	8.950	3.350	5.476
CCC	Speech	0.879	0.932	0.905	0.860	0.681	0.068	10.082	2.925	5.430
AAA	Speech	0.879	0.924	0.901	0.855	0.685	0.076	9.046	2.930	5.149
BBB	Speech	0.890	0.905	0.898	0.852	0.716	0.095	7.567	3.185	4.909
FFF	Text	0.892	0.878	0.885	0.836	0.729	0.122	5.998	3.247	4.413
DDD	Text	0.885	0.886	0.886	0.837	0.713	0.114	6.280	3.087	4.403
EEE	Text	0.893	0.865	0.879	0.828	0.736	0.135	5.449	3.280	4.227
Baseline	Text	0.892	0.858	0.875	0.823	0.734	0.142	5.176	3.232	4.090
MMM	Text	0.891	0.851	0.871	0.819	0.736	0.149	4.953	3.229	3.999
KKK	Text	0.891	0.847	0.868	0.816	0.736	0.153	4.822	3.213	3.936
LLL	Text	0.890	0.843	0.866	0.813	0.736	0.157	4.697	3.198	3.876
JJJ	Text	0.659	0.900	0.761	0.649	0.236	0.100	2.356	1.177	1.665

Table 1: *Results for 16 anonymised submissions and two baseline systems. Entries in parentheses are "non-official". "Rec" = recogniser used ("Text" = data pre-recognised using baseline Kaldi recogniser, "Speech" = other recogniser), "Pr" = precision, "R" = recall, "F" = F-measure, "SA" = scoring accuracy, "RCR" = relative correct rejections, "RFR" = relative false rejections, "D" = D-measure, "D_A" = D-measure on accepts, "D_{full}" = geometrical mean of D and DA, "Baseline" = system with baseline Kaldi recogniser and baseline XML grammar; "BaselinePerfectRec" = system with input from transcriptions and baseline XML grammar.*

We now extend D to the symmetrical form used in the current (third) edition. First we define D_A to be the metric corresponding to D for accepts as opposed to rejects; it is the ratio of the relative correct acceptance rate (RC_A) to the relative false acceptance rate (RF_A). Thus we put $RC_A = \frac{C_A}{F_R+C_A}$ and $RF_A = \frac{F_A}{C_R+F_A}$, then put

$$D_A = \frac{RC_A}{RF_A} = \frac{C_A/(F_R+C_A)}{F_A/(C_R+F_A)} = \frac{C_A(C_R+F_A)}{F_A(F_R+C_A)}$$

As with D, high values of D_A are good.

Finally, we define D_{full} to be the geometric mean of D_A and D. It turns out that this has a simple and natural form:

$$D_{full} = \sqrt{D_A D} = \sqrt{\frac{C_A(C_R+F_A)}{F_A(F_R+C_A)} \frac{C_R(F_R+C_A)}{F_R(C_R+F_A)}}$$

$$= \sqrt{\frac{C_A C_R}{F_A F_R}}$$

When announcing the task, we said that entries would be scored using D_{full}, with a k value of 3 and the added requirement that at least 50% of all incorrect responses should be rejected, and at least 50% of all correct responses should be accepted. This is a reasonable condition for any system that might be useful in practice, and prevents gaming of the metric by focusing exclusively on one component and ignoring the other.

3. Results

We received a total of 16 entries, ten using data pre-processed through the baseline Kaldi recogniser and six using a custom recogniser; three entries were received after the deadline and are marked as "non-official". Table 1 summarises the results. Three points immediately stand out:

1. There has been nontrivial progress. The best value of D_{full} in the second edition of the task was 5.691; this has now improved to 6.342. Although the 2018 and 2019 test sets are different, the "Baseline" system, which competed unchanged in both editions, provides a comparison point. Since its D_{full} score is similar in the 2019 set (3.954 versus 4.090), it is reasonable to assume that the 2019 set is comparable, and hence that the improvement in the top score is meaningful.
2. D_{full} aligns well with conventional metrics. The winning entry also gets the best score on F and SA
3. All the top spots are filled by entries who supplied their own speech recognition. Also, BaselinePerfectRec is far better than any of the real entries, despite the fact that it uses a naïve grammar-based language processing method which consists simply of a table lookup. It is tempting to draw the conclusion that the largest gains for now are to be found in doing better speech recognition. Note also that most of the variation is in D as opposed to D_A. There is little difference here among the "Text" entries, which all used the baseline recogniser.

4. What makes items difficult to label?

The test set contained 740 "correct" examples (the gold standard says the system should accept) and 260 "incorrect" examples (the gold standard says the system should reject). Following the analysis in the second edition overview paper [9], we ordered both the correct and incorrect subsets of the test data by the number of entries supplying the wrong judgement, assuming that, at least to a first approximation, examples where many systems gave an incorrect judgement were hard to label correctly and examples where few systems gave an incorrect judgement were easy. The approximate distribution is indicated in the first two columns of Table 2, where we aggregate the

#Bad	#Utts	Incom	Pron	Stutt	xTalk	SNSN	Faint	OOV	d(gr)	d(tr)	#WE	WER
“Correct” examples (should accept)												
0-2	605	1.16	0.00	3.64	0.83	2.98	2.31	1.65	0.71	0.83	0.17	4.16
3-9	65	7.69	0.00	9.23	13.85	3.08	4.62	1.54	0.82	2.26	0.95	19.25
10-18	70	7.14	0.00	8.57	8.57	17.14	5.71	0.00	1.59	4.40	2.67	45.74
all	740	2.30	0.00	4.59	2.70	4.32	2.84	1.49	0.80	1.29	0.47	9.42
“Incorrect” examples (should reject)												
0-2	184	21.74	11.41	9.78	2.17	3.26	4.89	2.72	3.35	6.27	1.48	43.07
3-9	31	16.13	6.45	9.68	3.23	6.45	0.00	0.00	2.39	4.45	1.35	26.59
10-18	45	26.67	0.00	2.22	0.00	2.22	4.44	0.00	2.60	3.49	1.22	26.74
all	260	21.92	8.85	8.46	1.92	3.46	4.23	1.92	3.11	5.57	1.42	38.28
Ratio of “Incorrect” to “Correct” values												
all	—	9.53	(∞)	1.84	0.71	0.80	1.49	1.29	3.89	4.32	3.02	4.06

Table 2: *Possible indicators of difficulty, broken down by number of entries out of 18 assigning the wrong label. First two columns: “#Bad” = number of entries assigning wrong label. “#Utts” = number of examples in group. Middle seven columns: percentage of items displaying seven types of possible problems. “Incom” = at least one incomprehensible word, “Pron” = at least one mispronounced word, “Stutt” = stuttering etc, “xTalk” = crosstalk, “SNSN” = strong non-speech noise, “Faint” = low volume in speech, “OOV” = at least one out of vocabulary word. Final three columns: average per-utterance value for three metrics. “d(gr)” = word edit distance to closest in-grammar example; “d(tr)” = character edit distance to closest correct training example; “#WE/WER” = number of word errors/word error rate in recognition hypothesis from baseline recogniser.*

data by dividing the sets into three bands labelled “easy” (0–2 wrong judgements), “medium” (3–9 wrong judgements) and “hard” (10–18 wrong judgements). On the left hand side of the table, we give the frequencies of occurrence for the extra annotation features described in §2.1. As in last year’s paper, we also added the following automatically computed labels:

OOV The transcription contains an OOV word, i.e. a word not in the training data or grammar.

d(gr) Edit distance, in words, between the transcription and the closest in-coverage sentence in the grammar.

d(tr) Edit distance, in characters, between the transcription and the closest correct example in the training data.

Word errors The number of word errors in the 1-best recognition result produced by the baseline recogniser.

Table 2 shows the distribution of the metrics over the three different bands. We used Light’s κ to estimate inter-annotator agreement: on this metric, agreement was “substantial” ($\kappa > 0.6$) for “Stutt”, “moderate” ($0.6 > \kappa > 0.4$) for “Pron”, “xTalk” and “SNSN” and “fair” ($0.4 > \kappa > 0.2$) for “Incom”.

The actual results are reassuringly similar to those from the second edition (compare with Table 2 of [9]). For the “Correct” subset, we see most of the metrics increasing as we move from the “easy” band to the ”hard” one, suggesting that they are relevant to predicting the difficulty of the labelling task. An ANOVA test supports this intuition. Five of the factors (“Incom”, “xTalk”, “d(gr)”, “d(tr)” and “#WE/WER”) are significant at $p < 0.001$. Word error accounts for 27% of the variance, and the remaining factors account for an additional 21%, giving a total of 48%; the corresponding figures for the second edition were 42%, 15% and 57%.

As in last year’s results, there is a sharp contrast with the “Incorrect” set. The patterns here are much less obvious, which agrees with the ANOVA figures. The only features significant even at $p < 0.01$ are “d(gr)” and “d(tr)”, together predicting a meagre 6% of the variance, and the remaining features add no more than another 5%.

It is again easy to see why the “Incorrect” set is much harder, with nearly all the potentially disturbing factors registering higher values (cf. the last line of the table). In particular, WER is four times as high.

5. Summary and further directions

We have presented an overview of the third edition of the Shared Task. The results suggest that the task has been quite successful. In the first edition, the winning score was less than 5, using the rather artificial D metric, which only really measured behaviour on rejections. The second edition improved this to a D score of 19. In the third edition, we have moved to D_{full}, a realistic metric which addresses both accepts and rejects. The best result is now a D_{full} of 6.34, composed of a D of 11.3 and a D_A of 3.54. The disparity between the D and D_A scores underlines the fact that incorrect responses are much harder to process. The best entries are getting values of D over 10, but are finding it hard to push D_A much above 3. Rejections are now quite informative and the student can take them seriously, but it is still very challenging to make accepts reliable.

The data we collected during 2014/15 is now exhausted. It would, however, be feasible to perform a new data-collection round. We propose the following plan. We would first create an updated version of the original app using the framework we are currently developing under Geneva’s CALLector project (https://www.unige.ch/callector). The new version would have the same content as the old one, but we would arrange things so that the central accept/reject decision could be sent to an arbitrary web-service. Thus any group with an existing Shared Task entry would be able to run it live, as long as they were able to package up their entry as a web service. Kay Berkling’s group at the University of Karlsruhe, Germany, is already performing CALL evaluation exercises with a local school, and would be in a position to organise the necessary data collection. If there is sufficient interest in continuing the Spoken CALL Shared Task, we will make this rough plan more concrete and try to implement it later in 2019.

6. References

[1] 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.

[2] C. Baur, C. Chua, J. Gerlach, M. Rayner, M. Russell, H. Strik, and X. Wei, "Overview of the 2017 Spoken CALL Shared Task," in *Proceedings of the Seventh SLaTE Workshop*, Stockholm, Sweden, 2017.

[3] M. Qian, X. Wei, P. Jančovič, and M. Russell, "The University of Birmingham 2017 SLaTE CALL Shared Task systems," in *Proceedings of the Seventh SLaTE Workshop*, Stockholm, Sweden, 2017.

[4] A. Magooda and D. Litman, "Syntactic and semantic features for human like judgement in spoken call," in *Proceedings of the Seventh SLaTE Workshop*, Stockholm, Sweden, 2017.

[5] Y. R. Oh, H.-B. Jeon, H. J. Song, B. O. Kang, Y.-K. Lee, J.-G. Park, and Y.-K. Lee, "Deep-Learning based automatic spontaneous speech assessment in a data-driven approach for the 2017 SLaTE CALL Shared Challenge," in *Proceedings of the Seventh SLaTE Workshop*, Stockholm, Sweden, 2017.

[6] K. Evanini, M. Mulholland, E. Tsuprun, and Y. Qian, "Using an automated content scoring system for spoken CALL responses: The ETS submission for the Spoken CALL Challenge," in *Proceedings of the Seventh SLaTE Workshop*, Stockholm, Sweden, 2017.

[7] N. Axtmann, C. Mehret, and K. Berkling, "The CSU-K rule-based pipeline system for Spoken CALL Shared Task," in *Proceedings of the Seventh SLaTE Workshop*, Stockholm, Sweden, 2017.

[8] A. Caines, "Spoken CALL Shared Task system description," in *Proceedings of the Seventh SLaTE Workshop*, Stockholm, Sweden, 2017.

[9] C. Baur, A. Caines, C. Chua, J. Gerlach, M. Rayner, M. Qian, M. Russell, H. Strik, and X. Wei, "Overview of the 2018 Spoken CALL Shared Task," in *Proc. Interspeech 2018*, Hyderabad, India, 2018.

[10] D. Jülg, M. Kunstek, C. Freimoser, K. Berkling, and M. Qian, "The csu-k rule-based system for the 2nd edition spoken CALL shared task," in *Proc. Interspeech 2018*, Hyderabad, India, 2018.

[11] H. Nguyen, L. Chen, R. Prieto, C. Wang, and Y. Liu, "Liulishuo's system for the spoken CALL shared task 2018," in *Proc. Interspeech 2018*, Hyderabad, India, 2018.

[12] M. Ateeq, A. Hanani, and A. Qaroush, "An optimization based approach for solving spoken CALL shared task," in *Proc. Interspeech 2018*, Hyderabad, India, 2018.

[13] M. Qian, X. Wei, P. Jancovic, and M. Russell, "The university of birmingham 2018 spoken CALL shared task systems," in *Proc. Interspeech 2018*, Hyderabad, India, 2018.

[14] K. Evanini, M. Mulholland, R. Ubale, Y. Qian, R. Pugh, V. Ramanarayanan, and A. Cahill, "Improvements to an automated content scoring system for spoken CALL responses: The ETS submission to the second spoken CALL shared task," in *Proc. Interspeech 2018*, Hyderabad, India.

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

[16] M. Rayner, N. Tsourakis, C. Baur, P. Bouillon, and J. Gerlach, "CALL-SLT: A spoken CALL system based on grammar and speech recognition," *Linguistic Issues in Language Technology*, vol. 10, no. 2, 2014.

[17] 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.

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