Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 378–385
Marseille, 11–16 May 2020

TLT-school: a Corpus of Non Native Children Speech

R. Gretter, M. Matassoni, S. Bannò, D. Falavigna
Fondazione Bruno Kessler (FBK)
Trento, Italy
{gretter, matasso, sbanno, falavi}@fbk.eu

Abstract

This paper describes "TLT-school", a corpus of speech utterances collected in schools of northern Italy for assessing the performance of students learning both English and German. The corpus was recorded in the years 2017 and 2018 from students aged between nine and sixteen years, attending primary, middle and high school. All utterances have been scored, in terms of some predefined proficiency indicators, by human experts. In addition, most of utterances recorded in 2017 have been manually transcribed carefully. Guidelines and procedures used for manual transcriptions of utterances will be described in detail, as well as results achieved by means of an automatic speech recognition system developed by us. Part of the corpus is going to be freely distributed to scientific community particularly interested both in non-native speech recognition and automatic assessment of second language proficiency.

Keywords: Children speech, Non-native speech, Language learning

1 Introduction

A large set of both written and spoken data has been collected during the implementation of campaigns aimed at assessing the proficiency of Italian students learning both German and English. Part of these data has been included in a corpus, named "Trentino Language Testing" in schools (TLT-school), that will be described in the following. Design of the campaigns, involvement of schools, data collection and annotation by human experts were carried out and coordinated by IPRASE, a research institute dealing with education which is located in Trentino, North Italy.

All the collected sentences have been annotated by human experts in terms of some predefined "indicators" which, in turn, were used to assign the proficiency level to each student undertaking the assigned test. This level is expressed according to the well-known Common European Framework of Reference for Languages (Council of Europe, 2001) scale. The CEFR defines 6 levels of proficiency: A1 (beginner), A2, B1, B2, C1 and C2. The levels considered in the evaluation campaigns where the data have been collected are: A1 (primary school), A2 (secondary school) and B1 (high school).

The indicators measure the linguistic competence of test takers both in relation to the content (e.g. grammatical correctness, lexical richness, semantic coherence, etc.) and to the speaking capabilities (e.g. pronunciation, fluency, etc.). Refer to Section 2 for a description of the adopted indicators.

The learners are Italian students, between 9 and 16 years old. They took proficiency tests by answering question prompts provided in written form. The "TLT-school" corpus, that we are going to make publicly available, contains part of the spoken answers (together with the respective manual transcriptions) recorded during some of the above mentioned evaluation campaigns. We will release the written answers in future. Details and critical issues found during the acquisition of the answers of the test takers will be discussed in Section 2.

The tasks that can be addressed by using the corpus are very challenging and pose many problems, which have only partially been solved by the interested scientific community.

From the ASR perspective, major difficulties observed in this corpus are represented by: *a)* recognition of both child and non-native speech, i.e. Italian pupils speaking both English and German, *b)* presence of a large number of spontaneous speech phenomena (hesitations, false starts, fragments of words, etc.), *c)* presence of multiple languages (English, Italian and German words are frequently uttered in response to a single question), *d)* presence of a significant level of background noise due to the fact that the microphone remains open for a fixed time interval (e.g. 20 seconds - depending on the question), and *e)* presence of non-collaborative speakers (students often joke, laugh, speak softly, etc.). Refer to Section 2.3 for a detailed description of the collected spoken data set.

Furthermore, since the sets of data from which "TLT-school" was derived were primarily acquired for measuring proficiency of second language (L2) learners, it is quite obvious to exploit the corpus for automatic speech rating. To this purpose, one can try to develop automatic approaches to reliably estimate the above-mentioned indicators used by the human experts who scored the answers of the pupils (such an approach is described in (Gretter et al., 2019)). However, it has to be noted that scientific literature proposes to use several features and indicators for automatic speech scoring which are partly different from those adopted in "TLT-school" corpus (see below for a brief review of the literature). Hence, we believe that adding new annotations to the corpus, related to particular aspects of language proficiency, can stimulate research and experimentation in this area.

Finally, it is worth mentioning that also written responses of "TLT-school" corpus are characterised by a high level of noise due to: spelling errors, insertion of word fragments, presence of words belonging to multiple languages, presence of off-topic answers (e.g. containing jokes, comments not related to the questions, etc.). This set of text data will allow scientists to investigate both language and behaviour of pupils learning second languages at school. Written data are described in detail in Section 2.2

Relation to prior work. Scientific literature is rich in approaches for automated assessment of spoken language proficiency. Performance is directly dependent on ASR accuracy which, in turn, depends on the type of input, read or spontaneous, and on the speakers' age, adults or children; see (Eskenazi, 2009) for an overview of spoken language technology for education. A recent publication reporting an overview of state-of-the-art automated speech scoring technology as it is currently used at Educational Testing Service (ETS) can be found in (Zechner and Evanini, 2019).

In order to address automatic assessment of complex spoken tasks requiring more general communication capabilities from L2 learners, the speaking part of the Arizona English Language Learner Assessment (AZELLA), a large-scale test (Cheng et al., 2014) developed by Pearson, has been used for some researches (Angeliki and Cheng, 2014; Cheng et al., 2014). The work described in (Cheng et al., 2014) reports results achieved on $1,500$ spoken tests, each double graded by human professionals, from a variety of tasks.

A public set of spoken data has been recently distributed in a spoken CALL (Computer Assisted Language Learning) shared task[1] where Swiss students learning English had to answer to both written and spoken prompts. The goal of this challenge is to label students' spoken responses as "accept" or "reject". Refer to (Baur et al., 2018) for details of the challenge and of the associated data sets.

Many non-native speech corpora (mostly in English as target language) have been collected over the years. A list, though not recent, as well as a brief description of most of them can be found in (Raab et al., 2007). The same paper also gives information on how the data sets are distributed and can be accessed (many of them are available through both LDC[2] and ELDA[3] agencies). Some of the corpora also provide proficiency ratings to be used in CALL applications. Among them, we mention the ISLE corpus (Menzel et al., 2000), which also contains transcriptions at the phonetic level and was used in the experiments reported in (Gretter et al., 2019).

Note that all corpora mentioned in (Raab et al., 2007) come from adult speech while, to our knowledge, the access to publicly available non-native children's speech corpora, as well as of children's speech corpora in general, is still scarce. Specifically concerning non-native children's speech, we believe worth mentioning the following corpora. The PF-STAR corpus (Batliner et al., 2005) contains English utterances read by both Italian and German children, between 6 and 13 years old. The same corpus also contains utterances read by English children. The *ChildIt* corpus (Russell, 2007) contains English utterances (both read and imitated) by Italian children.

By distributing "TLT-school" corpus, we hope to help researchers to investigate novel approaches and models in the areas of both non-native and children's speech and to build related benchmarks.

[1] https://regulus.unige.ch/spokencallsharedtask_3rdedition/ for details.

[2] https://www.ldc.upenn.edu/

[3] http://www.elra.info/en/about/elda/

Table 1: Evaluation of L2 linguistic competences in Trentino: level, grade, age and number of pupils participating in the evaluation campaigns. Most of the pupils took both the English and the German tests.

CEFR	Grade, School	Age	Number of pupils		
			2016	2017	2018
A1	5, primary	9-10	1074	320	517
A2	8, secondary	12-13	1521	111	614
B1	10, high school	14-15	378	124	1112
B1	11, high school	15-16	141	0	467
tot	5-11	9-16	3114	555	2710

Table 2: Written data collected during different evaluation campaigns. Column "#Q" indicates the total number of different (written) questions presented to the pupils.

Year	Lang	#Pupils	#Sentences	#Tokens	#Q
2016	ENG	3074	5062	299138	20
2016	GER	2870	4658	192144	25
2017	ENG	533	758	37225	5
2017	GER	529	745	30802	5
2018	ENG	2560	4600	293958	5
2018	GER	2200	3889	202309	5

2 Data Acquisition

In Trentino, an autonomous region in northern Italy, there is a series of evaluation campaigns underway for testing L2 linguistic competence of Italian students taking proficiency tests in both English and German. A set of three evaluation campaigns is underway, two having been completed in 2016 and 2018, and a final one scheduled in 2020. Note that the "TLT-school" corpus refers to only the 2018 campaign, that was split into two parts: 2017 try-out data set (involving about 500 pupils) and the actual 2018 data (about 2500 pupils). Each of the three campaigns (i.e. 2016, 2018 and 2020) involves about 3000 students ranging from 9 to 16 years old, belonging to four different school grade levels (5^{th}, 8^{th}, 10^{th}, 11^{th}) and three proficiency levels (A1, A2, B1). The schools involved in the evaluations are located in most part of the Trentino region, not only in its main towns; Table 1 highlights some information about the pupils that took part to the campaigns. Several tests, aimed at assessing the language proficiency of the students, were carried out by means of multiple-choice questions, which can be evaluated automatically. However, a detailed linguistic evaluation cannot be performed without allowing the students to express themselves in both written sentences and spoken utterances, which typically require the intervention of human experts to be scored.

Tables 2 and 3 report some statistics extracted from both the written and spoken data collected so far in all the campaigns. Each *written sentence* or *spoken utterance* received a total score by human experts, computed by summing up the scores related to 6 indicators in 2017/2018 (in the 2016 campaign, the number of indicators ranged from 3 to 6, according to the proficiency levels and the type of test). Each

Table 3: Spoken data collected during different evaluation campaigns. Column "#Q" indicates the total number of different (written) questions presented to the pupils. Column "Duration" relates to the total duration of the recordings of each row.

Year	Lang	#Pupils	#Utterances	Duration	#Q
2016	ENG	2748	17462	69:03:37	85
2016	GER	2542	15866	60:03:01	101
2017	ENG	511	4112	16:25:45	24
2017	GER	478	3739	15:33:06	23
2018	ENG	2332	15770	93:14:53	24
2018	GER	2072	13658	95:54:56	23

Table 4: List of the indicators used by human experts to evaluate specific linguistic competences.

lexical richness
pronunciation and fluency
syntactical correctness: morpho-syntactical correctness, orthography and punctuation
fulfillment on delivery: relevancy of the answer with respect to the prompt
coherence and cohesion
communicative, descriptive, narrative skills

indicator can assume a value 0, 1, 2, corresponding to bad, medium, good, respectively.
The list of the indicators used by the experts to score written sentences and spoken utterances in the evaluations, grouped by similarity, is reported in Table 4. Since every utterance was scored by only one expert, it was not possible to evaluate any kind of agreement among experts. For future evaluations, more experts are expected to provide independent scoring on same data sets, so this kind of evaluation will be possible.

2.1 Prompts

The speaking part of the proficiency tests in 2017/2018 consists of 47 question prompts provided in written form: 24 in English and 23 in German, divided according to CEFR levels. Apart from A1 level, which differs in the number of questions (11 for English; 10 for German), both English and German have 6 questions for A2 level and 7 questions for B1 level. As for A1 level, the first four introductory questions are the same (*How old are you?*, *Where do you live?*, *What are your hobbies?*, *Wie alt bist du?*, *Wo wohnst du?*, *Was sind deine Hobbys?*) or slightly different (*What's your favourite pet?*, *Welche Tiere magst du?*) in both languages, whereas the second part of the test puts the test-takers in the role of a customer in a pizzeria (English) or in a café (German). The presence of small differences between the questions in English and those in German is motivated by the need to make the tests less redundant and more engaging for students.
A2 level test is composed of small talk questions which relate to everyday life situations. In this case, questions are more open-ended than the aforementioned ones and allow the test-takers to interact by means of a broader range of answers. Finally, as for B1 level, questions are similar to A2 ones, but they include a role-play activity in the final part, which allows a good amount of freedom and creativity in answering the questions.

2.2 Written Data

Table 2 reports some statistics extracted from the written data collected so far. In this table, the number of pupils taking part in the English and German evaluation is reported, along with the number of sentences and tokens, identified as character sequences bounded by spaces.
It is worth mentioning that the collected texts contain a large quantity of errors of several types: orthographic, syntactic, code-switched words (i.e. words not in the required language), jokes, etc. Hence, the original written sentences have been processed to produce "cleaner" versions, in order to make the data usable for some research purposes (e.g. to train language models, to extract features for proficiency assessment, etc.).
To do this, we have applied some text processing, that in sequence:
• removes strange characters;
• performs some text normalisation (lowercase, umlaut, numbers, . . .) and tokenisation (punctuation, etc.)
• removes / corrects non words (e.g. *hallooooooooooo* becomes *hallo*; *aaaaaaaaeeeeeeeeiiiiiiii* is removed)
• identifies the language of each word, choosing among Italian, English, German;
• corrects common typing errors (e.g. *ai em* becomes *i am*)
• replaces unknown words, with respect to a large lexicon, with the label *<unk>*.
Table 5 reports some samples of written answers.

2.3 Spoken Data

Table 3 reports some statistics extracted from the acquired spoken data. Normally, around 20 students of the same class took the test together, at the same time and in the same classrooms, so it is quite common that speech of mates or teachers overlaps with the speech of the student speaking in her/his microphone. Also, the type of microphone depends on the equipment of the school. On average, the audio signal quality is nearly good, while the main problem is caused by a high percentage of extraneous speech. This is due to the fact that organisers decided to use a fixed duration - which depends on the question - for recording spoken utterances, so that all the recordings for a given question have the same length. However, while it is rare that a speaker has not enough time to answer, it is quite common that, especially after the end of the utterance, some other speech (e.g. comments, jokes with mates, indications from the teachers, etc.) is captured. In addition, background noise is often present due to several sources (doors, steps, keyboard typing, background voices, street noises if the windows are open, etc). Finally, it has to be pointed out that many answers are whispered and difficult to understand.

3 Manual Transcriptions

In order to create both an adaptation and an evaluation set for ASR, we manually transcribed part of the 2017 data

Table 5: Samples of written answers to English questions. On each line the CEFR proficiency level, the question and the answer are reported. Other information (session and question ID, total and individual scores, school/class/student anonymous ID) is also available but not included below.

CEFR	Question	Answer
A1	You are on a trip to Trentino with your family. Add a message to a picture you took and want to send it to a friend. Tell us: 1. where you are; 2.what you do; 3. what you like or dislike.	dear tiago . i'm swimming in the lake . there are some beautiful mountains . i swim in the levico lake . later i go home on the bikeand i eats ice creamwith my brother and my father . i like water is beautiful but i don't like the sun is very very hot ! is very impressive levico lake . see you soon . byee ! kacper ——— hello , i'm in the lake with my family . i play football with my dad and i eat a ice cream . the water is beautiful . it's very sunny . goodbye see you soon .
A2	Reply to Susan Hi! How are you? I've just received a new tennis racket. Would you like to meet at the sports centre and play a little? We can play for an hour and then we can get an ice cream together. Can you come at 5 o'clock? Don't forget to bring your tennis shoes, ok? I'm really looking forward to playing with you! Bye, Susan.	hello susan ! i'm fine thanks and you ? i'm very happy for you and for your message and i would like see your new racket but unfortunately today i can't come . tomorrow i have a very important football match and i must wake up at six o'clock so i need to sleep more than usually . we can meet ——— hello susan i'm fine . i'm sorry but i can't come with you , because i go to land between london for one concert at five o'clock . bye .
B1	Write an English post for your blog where you talk about what you need to do to learn a language well.	if you want to learn a new language you can make it . the first thing is studying many words and make a course . the second thing is go out of your state and arrive at the state where speak the language that you want learn . the first days are more difficult , but then its more easy . you must study the grammar too .
B1	Write a short email to a friend of yours to tell him / her that you intend to start studying another foreign language and what the reasons are.	hi sophie ! how are you ? i am writing to you because i desire to say you that i will start to study a new langueges . why i decide it ? because i wont to live in spain in the future . what do you thing ? i wait your answer . with love ! bye !

sets. An initial set of guidelines for the annotation was defined and adopted by 5 researchers to manually transcribe about 20 minutes of these audio data. This experience led to a discussion, from which a second set of guidelines originated, aiming at reaching a reasonable trade-off between transcription accuracy and speed. Briefly, the most important guidelines are:

- only the main speaker has to be transcribed; presence of other voices (schoolmates, teacher) should be reported only with the label "@voices",
- presence of whispered speech was found to be significant, so it should be explicitly marked with the label "()",
- badly pronounced words have to be marked by a "#" sign, without trying to phonetically transcribe the pronounced sounds; "#*" marks incomprehensible speech;
- speech in a different language from the target language has to be reported by means of an explicit marker *"I am 10 years old @it(io ho già risposto)"*.

Next, we concatenated utterances to be transcribed into blocks of about 5 minutes each. We noticed that knowing the question and hearing several answers could be of great help for transcribing some poorly pronounced words or phrases. Therefore, each block contains only answers to the same question, explicitly reported at the beginning of the block.

Table 6: Inter-annotator agreement between pairs of students in terms of words. Students transcribed English utterances first and German ones later.

High school	Language	#Transcribed words	#Different words	Agreement
C	English	965	237	75.44%
C	German	822	139	83.09%
S	English	1370	302	77.96%
S	German	1290	226	82.48%

Table 7: Statistics from the spoken data sets (2017) used for ASR.

ID	# of utt.	duration total	duration avg	tokens total	tokens avg
Ger Train All	1448	04:47:45	11.92	9878	6.82
Ger Train Clean	589	01:37:59	9.98	2317	3.93
Eng Train All	2301	09:03:30	14.17	26090	11.34
Eng Train Clean	916	02:45:42	10.85	6249	6.82
Ger Test All	671	02:19:10	12.44	5334	7.95
Ger Test Clean	260	00:43:25	10.02	1163	4.47
Eng Test All	1142	04:29:43	14.17	13244	11.60
Eng Test Clean	423	01:17:02	10.93	3404	8.05

At this point, we engaged about 30 students from two Italian linguistic high schools (namely "C" and "S") to perform manual transcriptions of most of the 2017 audio data.

After a joint training session, where the guidelines were explained and motivated, we paired students together. Each pair first transcribed, individually, the same block of 5 minutes. Then, they went through a comparison phase, where each pair of students discussed their choices and agreed on a single transcription for the assigned data. Transcriptions made before the comparison phase were retained to evaluate inter-annotator agreement. Apart from this first 5 minute block, each utterance was transcribed by only one transcriber. Inter-annotator agreement for the 5-minute blocks is shown in Table 6 in terms of words (after removing hesitations and other labels related to background voices and noises, etc.). The low level of agreement reflects the difficulty of the task.

In order to assure quality of the manual transcriptions, every sentence transcribed by the high school students was automatically processed to find out possible formal errors, and manually validated by researchers in our lab.

Speakers were assigned either to training or evaluation sets, with proportions of $\frac{2}{3}$ and $\frac{1}{3}$, respectively; then training and evaluation lists of utterances were built, accordingly. Table 7 reports statistics from the spoken data set. The ID *All* identifies the whole data set, while *Clean* defines the subset in which sentences containing background voices, incomprehensible speech and word fragments were excluded.

4 Usage of the Data

From the above description it appears that the corpus can be effectively used in many research directions.

4.1 ASR-related Challenges

The spoken corpus features non-native speech recordings in real classrooms and, consequently, peculiar phenomena appear and can be investigated. Phonological and cross-language interference requires specific approaches for accurate acoustic modelling. Moreover, for coping with cross-language interference it is important to consider alternative ways to represent specific words (e.g. words of two languages with the same graphemic representation).

Table 8, extracted from (Gretter et al., 2019), reports WERs obtained on evaluation data sets with a strongly adapted ASR, demonstrating the difficulty of the related speech recognition task for both languages. Refer to (Matassoni et al., 2018) for comparisons with a different non-native children speech data set and to scientific literature (Wilpon and Jacobsen, 1996; Das et al., 1998; Li and Russell, 2001; Giuliani and Gerosa, 2003; Potamianos and Narayanan, 2003; Gerosa et al., 2007; Gerosa et al., 2009; Liao et al., 2015; Serizel and Giuliani, 2016) for detailed descriptions of children speech recognition and related issues. Important, although not exhaustive of the topic, references on non-native speech recognition can be found in (Wang and Schultz, 2003; Wang et al., 2003; Oh et al., 2006; Strik et al., 2009; Steidl et al., 2004; Bouselmi et al., 2006; Duan et al., 2017; Li et al., 2016; Lee and Glass, 2015; Das and Hasegawa-Johnson, 2015).

As for language models, accurate transcriptions of spoken responses demand for models able to cope with not well-formed expressions (due to students' grammatical errors).

Table 8: WER results on 2017 spoken test sets.

German	English
42.6	35.9

Also the presence of code-switched words, words fragments and spontaneous speech phenomena requires specific investigations to reduce their impact on the final performance.

We believe that the particular domain and set of data pave the way to investigate into various ASR topics, such as: non-native speech, children speech, spontaneous speech, code-switching, multiple pronunciation, etc.

4.2 Data Annotation

The corpus has been (partly) annotated using the guidelines presented in Section 3 on the basis of a preliminary analysis of the most common acoustic phenomena appearing in the data sets.

Additional annotations could be included to address topics related to other spurious segments, as for example: understandable words pronounced in other languages or by other students, detection of phonological interference, detection of spontaneous speech phenomena, detection of overlapped speech, etc. In order to measure specific proficiency indicators, e.g. related to pronunciation and fluency, suprasegmental annotations can be also inserted in the corpus.

4.3 Proficiency Assessment of L2 Learners

The corpus is a valuable resource for training and evaluating a scoring classifier based on different approaches. Preliminary results (Gretter et al., 2019) show that the usage of suitable linguistic features mainly based on statistical language models allow to predict the scores assigned by the human experts.

The evaluation campaign has been conceived to verify the expected proficiency level according to class grade; as a result, although the proposed test cannot be used to assign a precise score to a given student, it allows to study typical error patterns according to age and level of the students.

Furthermore, the fine-grained annotation, at sentence level, of the indicators described above is particularly suitable for creating a test bed for approaches based on "word embeddings" (Chen et al., 2018; Oh et al., 2017; Qian et al., 2019) to automatically estimate the language learner proficiency. Actually, the experiments reported in (Chen et al., 2018) demonstrate superior performance of word-embeddings for speech scoring with respect to the well known (feature-based) SpeechRater system (Zechner et al., 2009; Zechner and Evanini, 2019). In this regard, we believe that additional, specific annotations can be developed and included in the "TLT-school" corpus.

4.4 Modelling Pronunciation

By looking at the manual transcriptions, it is straightforward to detect the most problematic words, i.e. frequently occurring words, which were often marked as mispronounced (preceded by label "#"). This allows to prepare

Table 9: Words suitable for pronunciation analysis. Data come from the 2017 manually transcribed data. Numbers indicate the number of occurrences, divided into test and training, with good and bad pronounciations.

word	tot occ	good vs. bad		good vs. bad	
German		Test GER		Train GER	
ich	1132	317	32	735	48
lieblingsessen	113	22	15	45	31
ist	374	131	9	211	23
mein	204	52	9	129	14
tschüss	33	6	5	7	15
höre	19	2	4	4	9
alt	191	67	7	111	6
frühstück	31	4	3	15	9
milch	22	1	6	9	6
heisse	54	14	3	30	7
English		Test ENG		Train ENG	
favourite	578	171	17	362	28
pet	169	49	10	96	14
thank	179	57	4	102	16
live	291	87	8	185	11
volleyball	97	22	5	60	10
football	246	75	7	157	7
years	170	47	2	109	12
subject	60	13	7	34	6
prefer	116	37	6	66	7
friends	120	27	3	82	8

a set of data composed by good pronounced vs. bad pronounced words.
A list of words, partly mispronounced, is shown in Table 9, from which one can try to model typical pronunciation errors (note that other occurrences of the selected words could be easily extracted from the non-annotated data). Finally, as mentioned above, further manual checking and annotation could be introduced to improve modelling of pronunciation errors.

5 Distribution of the Corpus

The corpus to be released is still under preparation, given the huge amount of spoken and written data; in particular, some checks are in progress in order to:

- remove from the data responses with personal or inadequate content (e.g. bad language);
- normalise the written responses (e.g. upper/lower case, punctuation, evident typos);
- normalise and verify the consistency of the transcription of spoken responses;
- check the available human scores and - if possible - merge or map the scores according to more general performance categories (e.g. delivery, language use, topic development) and an acknowledged scale (e.g. from 0 to 4)[4].

In particular, the proposal for an international challenge focused on non-native children speech recognition is being submitted where an English subset will be released and the perspective participants are invited to propose and evaluate state-of-art techniques for dealing with the multiple issues related to this challenging ASR scenario (acoustic and language models, non-native lexicon, noisy recordings, etc.).

6 Conclusions and Future Works

We have described "TLT-school", a corpus of both spoken and written answers collected during language evaluation campaigns carried out in schools of northern Italy. The procedure used for data acquisition and for their annotation in terms of proficiency indicators has been also reported. Part of the data has been manually transcribed according to some guidelines: this set of data is going to be made publicly available. With regard to data acquisition, some limitations of the corpus have been observed that might be easily overcome during next campaigns. Special attention should be paid to enhancing the elicitation techniques, starting from adjusting the questions presented to test-takers. Some of the question prompts show some lacks that can be filled in without major difficulty: on the one hand, in the spoken part, questions do not require test-takers to shift tense and some are too suggestive and close-ended; on the other hand, in the written part, some question prompts are presented both in source and target language, thus causing or encouraging code-mixing and negative transfer phenomena. The elicitation techniques in a broader sense will be object of revision (see (Cooke, 1994) and specifically on children speech (Beckman et al., 2017)) in order to maximise the quality of the corpus. As for proficiency indicators, one first step that could be taken in order to increase accuracy in the evaluation phase both for human and automatic scoring would be to divide the second indicator (pronunciation and fluency) into two different indicators, since fluent students might not necessarily have good pronunciation skills and vice versa, drawing for example on the IELTS [5] Speaking band descriptors. Also, next campaigns might consider an additional indicator specifically addressed to score prosody (in particular intonation and rhythm), especially for A2 and B1 level test-takers. Considering the scope of the evaluation campaign, it is important to be aware of the limitations of the associated data sets: proficiency levels limited to A1, B1 and B2 (CEFR); custom indicators conceived for expert evaluation (not particularly suitable for automated evaluation); limited amount of responses per speaker. Nevertheless, as already discussed, the fact that the TLT campaign was carried out in 2016 and 2018 in the whole Trentino region makes the corpus a valuable linguistic resource for a number of studies associated to second language acquisition and evaluation. In particular, besides the already introduced proposal for an ASR challenge in 2020, other initiatives for the international community can be envisaged: a study of a fully-automated evaluation procedure without the need of experts' supervision; the investigation of end-to-end classifiers that directly use the spoken response as input and produce proficiency scores according to suitable rubrics.

[4] https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf

[5] https://www.ielts.org

7 Acknowledgements

This work has been partially funded by IPRASE (http://www.iprase.tn.it) under the project "TLT - Trentino Language Testing 2018". We thank ISIT (http://www.isit.tn.it) for having provided the data and the reference scores.

8 Bibliographical References

Angeliki, M. and Cheng, J. (2014). Using Deep Neural Networks to improve proficiency assessment for children English language learners. In *Proc. of Interspeech*, pages 1468–1472.

Batliner, A., Blomberg, M., D'Arcy, S., Elenius, D., Giuliani, D., Gerosa, M., Hacker, C., Russell, M., Steidl, S., and Wong, M. (2005). The PF-STAR children's speech corpus. In *Proc. of Eurospeech*, pages 2761–2764.

Baur, C., Caines, A., Chua, C., Gerlach, J., Qian, M., Rayner, M., Russell, M., Strik, H., and Wei, X. (2018). Overview of the 2018 spoken call shared task. In *Proc. of Interspeech*, pages 2354–2358, Hyderabad, India.

Beckman, M., Plummer, A., Munson, B., and Reidy, P. F. (2017). Methods for eliciting, annotating, and analyzing databases for childspeech development. *Computer Speech Language*, (45):278–299.

Bouselmi, G., Fohr, D., Illina, I., and Haton, J. P. (2006). Multilingual non-native speech recognition using phonetic confusion-based acoustic model modification and graphemic constraints. In *Proc. of ICSLP*, pages 109–112.

Chen, L., Tao, J., Ghaffarzadegan, S., and Qian, Y. (2018). End-to-end neural network based automated speech scoring. In *Proc. of ICASSP*, pages 6234–6238, Calgary, Canada.

Cheng, J., Zhao-D'Antilio, Y., Chen, X., and Bernstein, J. (2014). Automatic spoken assessment of young english language learners. In *Proc. of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*.

Cooke, N. J. (1994). Varieties of knowledge elicitation techniques. *International Journal on Human-Computer Studies*, (41):801–849.

Das, A. and Hasegawa-Johnson, M. (2015). Cross-lingual transfer learning during supervised training in low resource scenarios. *Proc. of Interspeech*, pages 3531–3535.

Das, S., Nix, D., and Picheny, M. (1998). Improvements in Children's Speech Recognition Performance. pages 433–436, Seattle,WA, May.

Duan, R., Kawahara, T., Dantsuji, M., and Zhan, J. (2017). Articulatory modeling for pronunciation error detection without non-native training data based on dnn transfer learning. *IEICE Transactions on Information and Systems*, E100.D(9):2174–2182.

Eskenazi, M. (2009). An overview of spoken language technology for education. speech communication. *Speech Communication*, 51(10):2862–2873.

Gerosa, M., Giuliani, D., and Brugnara, F. (2007). Acoustic variability and automatic recognition of childrenâs speech. *Speech Communication*, 49(10â11):847 – 860.

Gerosa, M., Giuliani, D., and Brugnara, F. (2009). Towards age-independent acoustic modeling. *Speech Communication*, 51(6):499 – 509.

Giuliani, D. and Gerosa, M. (2003). Investigating Recognition of Children Speech. In *Proc. of ICASSP*, volume 2, pages 137–40, Hong Kong, Apr.

Gretter, R., Matassoni, M., Allgaier, K., Tchistiakova, S., and Falavigna, D. (2019). Automatic assessment of spoken language proficiency of non-native children. In *Proc. of ICASSP*.

Lee, A. and Glass, J. (2015). Mispronunciation detection without nonnative training data. In *Proc. of Interspeech*, pages 643–647.

Li, Q. and Russell, M. (2001). Why is Automatic Recognition of Children's Speech Difficult?". In *Proc. of Eurospeech*, Aalborg, Denmark, Sept.

Li, W., Siniscalchi, M., Chen, N. F., and Lee, C. H. (2016). Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling. In *Proc. of ICASSP*, pages 6135–6139.

Liao, H., Pundak, G., Siohan, O., Carroll, M., Coccaro, N., Jiang, Q., Sainath, T. N., Senior, A., Beaufays, F., and Bacchiani, M. (2015). Large vocabulary automatic speech recognition for children. In *Proc. of Interspeech*.

Matassoni, Gretter, R., Falavigna, D., and Giuliani, D. (2018). Non-native children speech recognition through transfer learning. In *Proc. of ICASSP*.

Menzel, W., Atwell, E., Bonaventura, P., Herron, D., Howarth, P., Morton, R., and Souter, C. (2000). The ISLE corpus of non-native spoken English. In *Proc. of LREC*, pages 957–964.

Oh, Y. R., Yoon, J. S., and Kim, H. K. (2006). Adaptation based on pronunciation variability analysis for non native speech recognition. In *Proc. of ICASSP*, pages 137–140.

Oh, Y. R., Jeon, H.-B., Song, H. J., Kang, B. O., Lee, Y.-K., Park, J., and Lee, Y.-K. (2017). Deep-learning based Automatic Spontaneous Speech Assessment in a Data-Driven Approach for the 2017 SLaTE CALL Shared Challenge. In *Proc. of SlaTe*, pages 103–108, Stockholm, Sweden.

Potamianos, A. and Narayanan, S. (2003). Robust Recognition of Children's Speech. 11(6):603–615, Nov.

Qian, M., Jancovic, P., and Russel, M. (2019). The university of birmingham 2019 spoken call shared task systems: Exploring the importance of word order in text processing. In *Proc. of SlaTe*, pages 11–15, Graz, Austria.

Raab, M., Gruhn, R., and Noeth, E. (2007). Non-native speech databases. In *Proc. of ASRU*, pages 413–418, Kyoto, Japan.

Russell, M. (2007). Analysis of Italian children's English pronunciation. *http://archive.is/http://www.eee.bham.ac.uk/russellm*.

Serizel, R. and Giuliani, D. (2016). Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children. *Natural Language Engineering*, FirstView:1–26, 7.

Steidl, S., Stemmer, G., Hacker, C., and Nöth, E. (2004).

Adaptation in the pronunciation space for non-native speech recognition. In *Proc. of ICSLP*, pages 2901–2904.

Strik, H., Truong, K., de Wet, F., and Cucchiarini, C. (2009). Comparing different approaches for automatic pronunciation error detection. *Speech Communication*, 51(10):845–852.

Wang, Z. and Schultz, T. (2003). Non-native spontaneous speech recognition through polyphone decision tree specialization. In *Proc. of Eurospeech*, pages 1449–1452.

Wang, Z., Schultz, T., and Waibel, A. (2003). Comparison of acoustic model adaptation techniques on non-native speech. In *Proc. of ICASSP*, pages 540–543.

Wilpon, J. and Jacobsen, C. (1996). A Study of Speech Recognition for Children and Elderly. pages I–349–352, Atlanta, GA, May.

Zechner, K. and Evanini, K. (2019). *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech*. Educational Testing Service, Princeton (NJ).

Zechner, K., Higgins, D., Xi, X., and Williamson, D. (2009). Automatic scoring of nonânative spontaneous speech in tests of spoken English. *Speech Communication*, 51(10):883–895.