Collecting Linguistic Resources for Assessing Children's Pronunciation of Nordic Languages

Anne Marte Haug Olstad[1,*], **Anna Smolander**[2,*], **Sofia Strömbergsson**[3], **Sari Ylinen**[2], **Minna Lehtonen**[4], **Mikko Kurimo**[5], **Yaroslav Getman**[5], **Támas Grosz**[5], **Xinwei Cao**[6], **Torbjørn Svendsen**[6], **Giampiero Salvi**[6]

[1] University of Oslo, Norway, a.m.h.olstad@iln.uio.no
[2] Tampere University, Finland, {anna.2.smolander, sari.p.ylinen}@tuni.fi
[3] Karolinska Institutet, Sweden, sofia.strombergsson@ki.se
[4] University of Turku, Finland, minna.h.lehtonen@utu.fi
[5] Aalto University, Finland, {mikko.kurimo, yaroslav.getman, tamas.grosz}@aalto.fi
[6] NTNU, Norway, {xinwei.cao, torbjorn.svendsen, giampiero.salvi}@ntnu.no
* Sharing first authorship.

Abstract

This paper reports on the experience collecting a number of corpora of Nordic languages spoken by children. The aim of the data collection is providing annotated data to develop and evaluate computer assisted pronunciation assessment systems both for non-native children learning a Nordic language (L2) and for L1 children with speech sound disorder (SSD). The paper presents the challenges encountered recording and annotating data for Finnish, Swedish and Norwegian, as well as the ethical considerations related with making this data publicly available. We hope that sharing this experience will encourage others to collect similar data for other languages. Of the different data collections, we were able to make the Norwegian corpus publicly available in the hope that it will serve as a reference in pronunciation assessment research.

Keywords: pronunciation assessment, CAPT, child speech, second language acquisition, speech sound disorder, Nordic languages

1. Introduction

It requires a lot of practice and accurate feedback to properly learn the pronunciation of a foreign or second language (L2). The rapid advance of speech and language technology has created automated tools that enable extensive self-practice when human teachers are not available. However, most effort in speech technology development is directed to the most popular languages and adult learners. This is motivated by commercial reasons and the availability of large quantities of annotated speech data required to train the large speech and language models for automatic speech recognition (ASR) and automatic pronunciation assessment (APA). Most openly available speech data are recorded from native and fluent adult speakers, but language learners' and particularly children's speech have very different acoustic and linguistic characteristics that are poorly covered in native and adult speech. The situation is worse for child learners in low-resource target languages, where no open speech data are available.

Corpora suitable for research on child speech recognition exist for several languages. Examples include American English, e.g.; CMU Kids (Maxine Eskenazi and Graff, 1997), CSLU Kids (Khaldoun Shobaki and Cole, 1997), My Science Tutor (Sameer Pradhan and Ward, 2021); Dutch (C. Cucchiarini and Smits, 2008); Mandarin (Yu et al., 2021) and German (Rumberg et al., 2022). These data sets, however, do not include recordings of L2 learners, and do not provide pronunciation scores. Data sets for evaluating pronunciation scoring exist for adult speakers, e.g.; for non-native speakers of English (Zhang et al., 2021), (Vidal et al., 2019), (Sudhakara et al., 2019). A smaller data set for assessing children's pronunciation of English as second language was created by Shi et al. (2020). However, most of the data had to be later erased because of privacy. Finally, there are to the authors' knowledge no corpora that address non-standard or L2 pronunciation of Nordic languages for child speakers.

In this paper, we describe our recent efforts and experiences in collecting and releasing children's speech data for Norwegian, Swedish and Finnish and training and finetuning corresponding large speech and language models for ASR and APA development. We also describe how we used these models to develop baseline ASR and APA systems for interactive mobile game applications to aid pronunciation learning in young children learning L2 and children with speech sound disorders (SSD) (Getman et al., 2023b).

One particularly important step after recording the child speech data was to manually rate the pronunciation in all samples. Because we wanted to train an APA system to provide a quick and easily understandable rating scale for the children playing the pronunciation learning game, we first needed to define such a rating scale for the human raters to reliably and consistently annotate the pronunciation of every word in the training data.

The main contributions of this paper are:

1. Sharing our experiences and best practices to collect, annotate and release child speech data for L2 and SSD in low-resource target languages.
2. Sharing the child speech data for L2 Norwegian.
3. Evaluating the ASR and APA performance and sharing the baseline systems that we obtained using this open data.

2. Data Collection

All the corpora of speech resources described here, with the exception of SweSSD described in Section 4, were collected with the goal of developing ASR and APA models that could be used in a digital language learning game for children. The Pop2Talk game (Karhila et al., 2017) aims to provide a fun way for children to practice pronunciation in a foreign language. In Pop2Talk, the players hear spoken words in the target language, and are then prompted to repeat the word they have heard. They then hear their own utterance played back to them and receive 1 to 5 stars generated by the automatic speech recognizer, based on how accurate it deems their pronunciation to be.

The method for collecting these corpora of speech data was driven by the Pop2Talk application and was very similar across languages and speakers (except SweSSD). First, a native "model" speaker of the target language was recorded saying all the relevant words. An audio track was then prepared using Audacity[1], containing all the model speaker words separated by silent gaps of ∼3 seconds. Then a label-track was added, and all the silent breaks were labeled with the target word. This Audacity project was used as a recording template for all the participants.

The instructions to the participants were to wear a headset with a microphone, listen to the words, and try to repeat what they heard as best as they could. The Audacity project template was used to reproduce the target words at the same time as recording the participant. The participant's utterances would fill the empty gaps that were previously labeled. After completion of the recording session, the Audacity function of "export multiple tracks" was used to export all the words uttered by the participant as individual tracks.

Score	Label
1	Not at all identifiable as the target word
2	Difficult to identify as the target word
3	Slight phonemic error(s)
4	Subphonemic error(s) or ”unexpected variants”
5	Prototypical, adult-like

Table 1: Labels guiding the evaluation of speech samples with reference to the global 1-5 scale.

Although this procedure does not allow for randomization of the word order, it was deemed to be acceptable given that it simplified the data collection significantly and allowed us to record a larger number of participants with the same allocated resources.

Some of the data sets were collected in-person with the participant and the experimenter sitting together in a quiet room (TeflonNorL2), and some (TeflonSweL2 and TeflonFinL2) were collected online using Zoom[2]. For TeflonSweSSD, both an in-person and a Zoom setting were used.

All the participants were children between 4 and 12 years old. Many children, especially the younger ones, required facilitation to keep their patience and concentration, as the whole process of hearing and repeating would typically take around 15 minutes without breaks. The children were instructed that they should let the experimenter know if a break was needed. The experimenter would also always pay close attention and suggest a break when it seemed necessary. Some children also liked having a soft toy animal in front of them, and act as if they were saying the words to the toy. Allowing them to play with modeling clay or another soft toy worked well for some children. Letting the children draw while speaking was attempted, but did not work well because of the noise generated by drawing.

3. Annotations

All the data were annotated according to general criteria agreed upon during the TEFLON project. Orthographic annotations are given for each utterance. Additionally, a global 1–5 score of speech accuracy is given for each utterance. The expertise of the evaluators varied between the different data sets, from speech-language pathologists, to linguists and phoneticians to master students specializing in language teaching. More details will be given in Section 4 for each corpus. To guide the evaluators in their use of the 1–5 scale, the scale steps were defined as described in Table 1. The scale steps were intended to present different levels of intelligibility. This 1–5 scale was the result of an iterative process involving growing experience with annotations. An initial 10-level scale was first

[1] https://audacityteam.org/

[2] https://zoom.us/

used. However, this scale was soon abandoned because of the need to adapt the definition to the different target languages and to achieve higher inter- and intra-rater agreement.

Some of the corpora were also augmented with extra annotations to afford specific tasks. Two examples of annotations are percentage of consonants correct (PCC), and percentage of phonemes correct (PPC) (Shriberg et al., 1997; Shriberg and Kwiatkowski, 1982). In some cases, a binary score was given to each phoneme in the canonical pronunciation of each word, where 1 corresponds to correctly pronounced and 0 to incorrectly pronounced. Deletions were marked in this case with a "-" symbol. In these annotations, however, there is no indication of what sounds were produced in place of the mispronounced phoneme. Also, binary annotations were given for noise/disruption, pre-speech noise, and repetition. These extra annotations had the effect of guiding the choice for the 1–5 scale, although the global assessment was not exclusively based on the accuracy of pronunciation of the single phonemes, as will be clear in Section 5. Moreover, these more specific annotations may be useful to the development of the automatic pronunciation assessment system if more detailed feedback compared to the global score is required. For more details, refer to Section 4.

4. Corpora

The data collection was carried out in the three Nordic languages, Swedish, Finnish and Norwegian. Details on each corpus are given here. A summary of some aspects of the data is given in Table 2.

4.1. Swedish

For the Swedish language, we collected three corpora. The first two, SweSSD and TeflonSweSSD were created at Karolinska Institutet and contain recordings of children speaking Swedish as their strongest language. The third corpus, TeflonSweL2 was created by Tampere University and contains recordings of Finnish children learning Swedish. These children had not studied Swedish at school before the data collection.

SweSSD: This dataset, briefly described in (Getman et al., 2023b), was collected within the project *Functional consequences of children's misarticulated speech* (Strömbergsson et al., 2020), based at Karolinska Institutet. The dataset included 6027 isolated word tokens (1109 unique words), recorded from 28 native Swedish children in the ages 4-10 years. To avoid identification of speakers (as the data was collected within a different project), all recordings were pseudonymized, and no information concerning the speakers was shared outside of Karolinska Institutet. During the TEFLON project, 1–5 scale scores were added to the corpus to make it homogeneous with the other TEFLON corpora.

TeflonSweSSD: This dataset was collected within TEFLON. As such, it was specifically tailored for the development of the game, intended for use with children with speech sound disorder (SSD), with specific difficulties with producing velar and/or fricative sounds. A target word list of 142 items was designed, with words containing velar and fricative sounds in initial, medial, and final word position. The target words were presented via headphones to the child speakers by a prerecorded adult speaker, and the task for the children was to repeat the word they heard. 35 children in the ages 4-8 years, with Swedish as their strongest language, participated as speakers. Information was not collected concerning whether the children had an SSD or not, but instead with the explicit goal of recording "different ways of speaking" (see also Table 2). As the speech error patterns in focus (velar fronting and stopping) for the intended speech game are found both in early typical speech development and later in children with protracted development, the information of whether a speaker had SSD or not was deemed irrelevant. 26 speakers were recorded in their homes, by a visiting project assistant, 8 speakers were recorded via Zoom, and one speaker was recorded in their home by their parent. Hence, recording conditions varied, but were considered ecologically valid in relation to the intended use of the game. The target words were then manually annotated and extracted from the full recordings in Audacity.

For the SweSSD and TeflonSweSSD datasets the orthographic transcription and the 1–5 global score of speech accuracy were created by speech-language pathologists according to the method described in Section 3. In addition, the same two evaluators scored the TeflonSweSSD recordings with reference to the Percentage of Consonants Correct (PCC), and to the Percentage of Phonemes Correct (PPC) (Shriberg et al., 1997; Shriberg and Kwiatkowski, 1982).

TeflonSweL2: The dataset was collected within TEFLON with the goal of aiding the development of the game. The intended use was for children learning Swedish as L2. The participants had Finnish as their strongest language, which guided the design of the corpus. The target word list contained 121 unique words that were considered important when learning elementary level Swedish, were not cognates of Finnish words, and also contained sounds not present in Finnish. The target words were recorded as described in Section 2. A toy animal was used as a proxy to which the child repeated the word. As the recording sessions

Corpus	lang.	speaker kind	# speakers (SSD or L2)	ages	# utterances (minutes)	# words	annotations	public availability
SweSSD	swe	Native/SSD	28 (16)	4–10	6027 (125)	1109	orth, glob	No
TeflonSweSSD	swe	Native/SSD	35 (NA*)	4–8	5012 (101)	142	orth, glob, PCC, PPC	No
TeflonSweL2	swe	L2	20 (20)	7–11	2384 (90)	121	orth, glob	No
TeflonFinL2	fin	L2	24 (24)	7–11	2124 (83)	90	orth, glob	No
TeflonNorL2	nor	Native/L2	52 (33)	5-12	9443 (544)	205	orth, glob, phon	Yes

Table 2: Summary of collected corpora. Legend: orth: orthographic, glob: global 1–5 score, PCC: percentage of consonants correct, PPC: percentage of phonemes correct, phon: binary score per phoneme. *) As explained in Section 4, TeflonSweSDD, this information was not included in the corpus by choice.

were relatively long from the kids' perspective, it was sometimes necessary to take small breaks every 3–4 minutes or change the toy animal, while some children were able to record all the words in one go. A quality check was implemented for each file by manually listening to them. 20 Finnish children of ages 7 to 11 who had not studied the target language at school yet participated as speakers. The collected data were rated by native Finnish university students completing the last year of their master's studies, majoring in Swedish language, specializing in language teaching, and with practical teaching experience. Furthermore, the annotators were trained as described in (Getman et al., 2023b). In total, the Swedish L2 data consists of 2384 speech utterances with total duration of 90 minutes (See Table 2 for details).

4.2. Finnish

The **TeflonFinL2** corpus was collected within TEFLON with the aim of developing a language learning game for children not native in the Finnish language. The word list contains 90 words that were considered essential in elementary Finnish, were not present in the children's native language or were expected to be difficult for L2 learners in general, such as Finnish words that contain front vowels ä, ö, and y (/æ/, /ø/, and /y/). The Finnish recordings were collected from 24 Ukrainian children aged 7 to 11 whose mother tongue was Ukrainian or Russian. The children had not learned the Finnish language at the time of the data collection. The data was recorded according to the procedure described in Section 2. A quality check was implemented for each file by manually listening to them. The collected data were rated and annotated according to the procedure described in Section 3. The data are composed of 2124 utterances with total duration of 83 minutes.

4.3. Norwegian

The **TeflonNorL2** corpus was collected within the TEFLON project according to the guidelines described in Section 2 and 3. The word list included 205 items containing different Norwegian speech sounds assumed to be difficult for beginner learners of Norwegian (e.g., Engen and Kulbrandstad, 2004; Hvenekilde, 1990). The data collection was conducted in two different periods. The first round of data collection involved 19 L1 Norwegian children in a school in eastern Norway and 13 children located in Finland with no previous exposure to Norwegian language. The latter had different L1s (Finnish, Estonian, Ukrainian, Russian). All children in this data collection were 5-10 years old. In the second data collection 20 children were recorded at a different school in eastern Norway following the same procedure. These children were 10-12-year-old L2 speakers and were all beginner learners of Norwegian. Their L1s include Dutch, Mandarin, English, Urdu, Vietnamese, Persian, Montenegrin, Ukrainian, Albanian and Russian. The children had lived in Norway for 1 to 10 months, with low enough Norwegian proficiency to have the right to extra Norwegian teaching ((Opplaeringslova, 1998) or (Kommune, 2023)).

All recording sessions were conducted in a quiet room, according to the procedure described in Section 2.

Speech assessments were performed on recordings from the second round of data collection, as well as on recordings of the overlapping items from the first round. The complete dataset thus contains recordings of 52 children (19 Norwegian and 33 non-Norwegian).

Two native Norwegian speakers with a background in linguistics provided the 1–5 scale assessment described in Section 3. In order to help with scoring 1 to 5, phonological transcriptions of every item in the word list were obtained using the NLB pronunciation lexicon for Norwegian Bokmål[3], and the two assessors marked every phoneme as correct or not. This extra scoring made it easier for two assessors to agree on a global score of each item, because it was to a degree possible to count the number of errors in an item.

About one third of all items were assessed by both assessors, and one third was assessed individually by each assessor. The assessors met regu-

[3] https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-52/

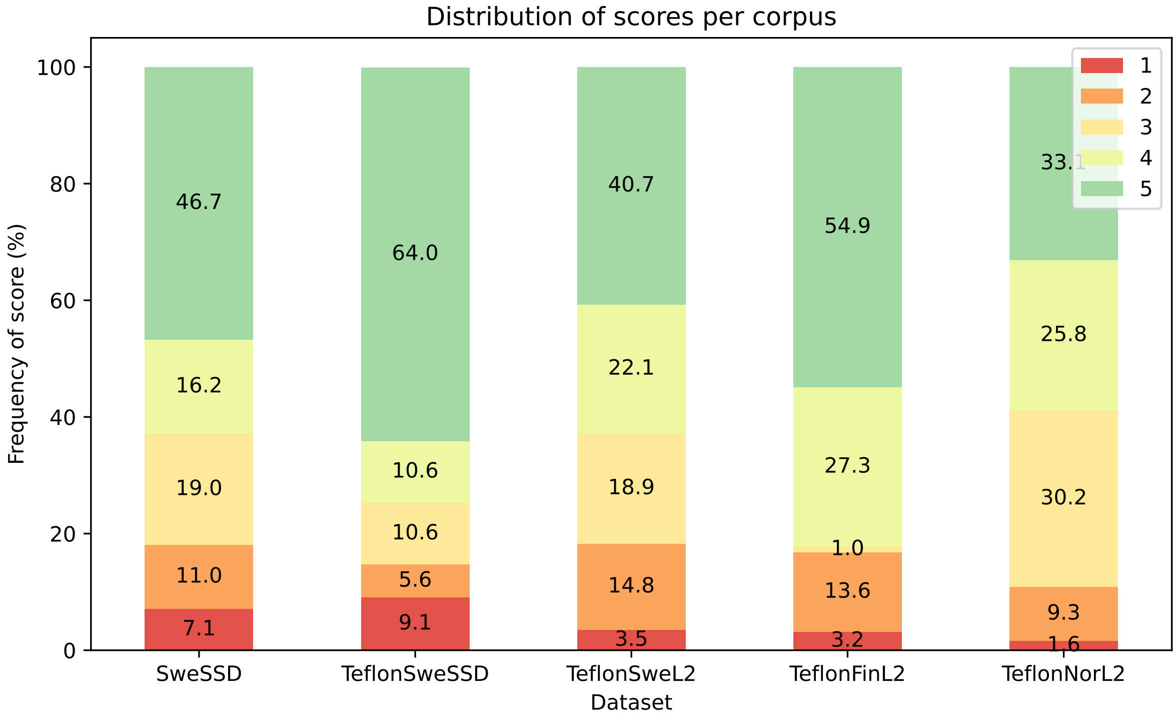

Figure 1: Score distribution for the different corpora.

larly and discussed difficult items, and the items assessed by both were distributed over time. The items assessed by both were first assessed by the two assessors separately and then the items where their scoring differed were discussed. The reason for difference in scores was typically either human error, using different equipment to listen to recordings, and sometimes disagreements about how severe the error was.

We make the corpus publicly available through the Norwegian Language Bank (Språkbanken)[4].

5. Statistics

The score distribution for the different corpora is displayed in Figure 1. As can be seen from the figure, score 4 and 5 are usually predominant and score 1 is the least represented in the data. This unavoidable class imbalance should be taken into account when building automatic pronunciation assessment systems.

In the following, we will give details on the TeflonNorL2 corpus which will we make publicly available. Figure 2 shows the distribution of pronunciation lengths (in terms of number of phonemes) for each utterance. Note that this measure is based on the canonical pronunciation of each word, given that the corpus does not include phonetic annotations. The majority of utterances has a length of 4 phonemes followed by 3 and 5. Only 3.8% of the utterances have length 7.

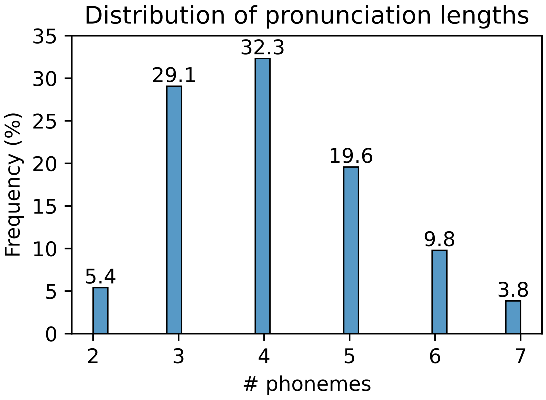

Figure 2: Distribution of pronunciation lengths (# of phonemes) per utterance in TeflonNorL2.

Figure 3 shows the distribution of phoneme errors in the data. For each utterance, the error is calculated as the percentage of phonemes that are incorrectly pronounced in that utterance. Circa 44% of the utterances have no phoneme errors. However, only 33.1% of utterances received the score 5 (see Figure 1). This means that, although the annotators did not judge any of the phonemes to be completely wrong, in some cases, they judged

[4] https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-94/, hdl:21.11146/94

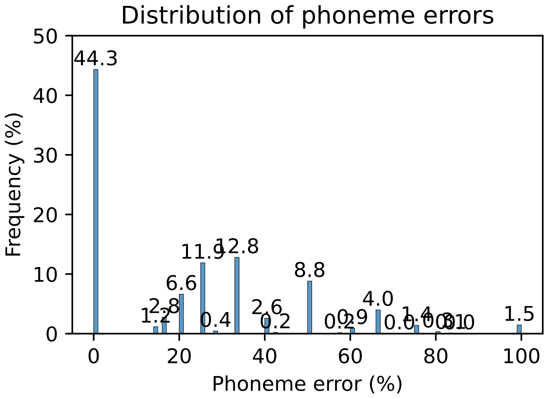

Figure 3: Distribution of pronunciation errors (# of wrong phonemes/pronunciation length) per utterance in TeflonNorL2.

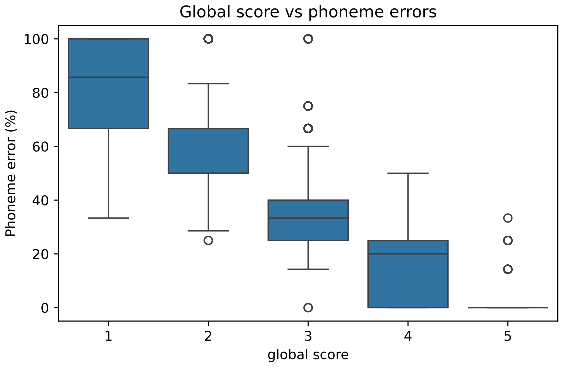

Figure 4: Global scores versus phoneme errors in TeflonNorL2.

the pronunciation as a whole not to be perfect.

Finally, Figure 4 shows the relationship between the percentage of phoneme errors and the global score. As espected, the figure shows a strong negative correlation between the two metrics (Pearson correlation coefficient -0.836). However, the ranges of phonemes errors that correspond to a certain global score are slightly overlapping (see the wiskers in the box plot). This is because the annotators, when assigning the global score did not only consider phonemic accuracy, but other global factors that are commonly associated with proficiency, such as prosody and hesitations.

For TeflonNorL2 about 29.8% of the examples were scored by two different annotators. For those examples, we computed inter-annotator agreement in three different ways. The percentage of times that the two annotators perfectly agreed was 75.3% (2122 cases out of 2818). In order to check how severe the disagreement was, we computed a confusion matrix that is given in Table 3. In the table, we can see that most of the confusions are between the higher levels in the scale and do not usually differ by more than two levels of the scale. Finally, we computed a Spearman correlation coefficient of 0.84 (p value=0.0) between the two sets of assessments.

	1	2	3	4	5
1	32	11	3	0	1
2	5	238	31	0	1
3	0	64	626	98	14
4	0	6	135	462	85
5	1	2	72	167	764

Table 3: Confusion matrix of annotations for 29.8% of the examples in TeflonNorL2 that were annotated twice.

Data	WER	CER	ACC	UAR	MAE
	[%] (↓)	[%] (↓)	[%] (↑)	[%] (↑)	(↓)
SweSSD*	17.17	6.42	60.31	47.64	.53
TeflonSweL2*	9.95	4.04	48.24	35.12	.70
TeflonFinL2*	6.30	2.13	72.08	43.07	.37
TeflonNorL2	10.74	4.21	55.18	39.83	.53

Table 4: Evaluation results of multi-task models trained on children speech datasets. The performance measures are explained in the text. (*) reproduced from (Getman et al., 2023b).

6. ASR and APA experiments

In this section, we provide preliminary experiments in using the corpora to train a system for ASR and APA. Experiments for SweSSD, TeflonSweL2, and TeflonFinL2 are reported from (Getman et al., 2023b), whereas TeflonNorL2 results are novel. For TeflonNorL2, we followed the same training setup and procedure proposed for SweSSD, TeflonSweL2, and TeflonFinL2 in (Getman et al., 2023b). For evaluation, we opted for 6-fold cross-validation (CV) with no fold or speaker overlap, resulting in 4 folds with 9 speakers and 2 folds with 8 speakers. The APA performance is measured with accuracy (ACC), unweighted average recall (UAR), and mean absolute error (MAE). Because the APA system uses ASR models, we also measured the ASR performance with word and character error rate (WER and CER).

We applied multi-task learning to optimize a single wav2vec 2.0 model (Baevski et al., 2020) simultaneously for ASR and APA. This means that the same model had two outputs, one for ASR (using connectionist temporal classification, CTC) and one for APA. The system was fine-tuned jointly with CTC and cross-entropy (CE) loss. As a foundation model, we used wav2vec 2.0 originally pretrained on Swedish and fine-tuned for Norwegian ASR by the AI-Lab at the National Library of Norway (De La Rosa et al., 2023).

Table 4 summarizes previous multi-task wav2vec 2.0 results from (Getman et al., 2023b) and the

System	WER [%] (↓)	CER [%] (↓)	ACC [%] (↑)	UAR [%] (↑)	MAE (↓)
CER_DT (baseline)	*N/A*	*N/A*	42.75	42.32	.87
MT_W2V2	10.74	4.21	55.18	39.83	.53
↳ + CER_DT			55.62	42.35	.53
MT_W2V2 (L20)	12.67	4.91	54.24	40.97	.57
↳ + CER_DT			55.39	44.04	.56

Table 5: Experiments on TeflonNorL2. The multi-task systems (MT_W2V2) can do both ASR and APA. The Transformer layer number preceding the classification head is in parenthesis if other than the last layer (L24).

new experiments with TeflonNorL2. However, the models trained for the different datasets are not fully comparable, because the words and the language are different as well as the amount of training data and the number of unique words. Also, the difficulty of pronouncing the words and the level of the speakers are hard to compare across the datasets.

We also run similar multi-task experiments as in (Getman et al., 2023a) by selecting the optimal hidden layer of wav2vec 2.0 to APA as well as combining wav2vec 2.0 with a simple decision tree (DT) trained on the CERs between the target words and the ASR outputs. The results in Table 5 confirm the previous findings that the last Transformer layer is not always the best for speech rating tasks and adding an external DT classifier can further improve the rating performance.

7. Making the data publicly available

The intention when collecting the data described in this paper was to make all the corpora publicly available. For the speech data collected in Sweden and Finland, the intended distribution channel woukd be through the respective nodes in the European Language Grid, that is, Språkbanken Tal[5] in Sweden and Kielipankki[6] in Finland. Unfortunately, however, with the current Swedish interpretation of the European legislation sharing of speech data is not allowed. Instead, by recommendation from Språkbanken Tal, we opted to include a description in the participant consent information, specifying that only *derivatives* from the speech data would be made publicly available.

In a similar vein, in Finland, children's speech data were interpreted as personal information that should not be publicly shared. This is because children's identity may potentially be recognized from the recordings, although the content of the recordings was not particularly sensitive (isolated words). These restrictions seem to be age dependent, as Finnish speech data from teenagers (over 16 years) could be shared in the past with consent from their parents.

In contrast to the other datasets, the Norwegian data in TeflonNorL2 was approved to be made publicly available (see Section 4.3 for the download link).

The different responses to our requests to the ethical committees in the different countries illustrate the complexity of speech data sharing. It is useful to stress that the L2 data sets for Swedish, Norwegian and Finnish were completely equivalent in all respects: age of participants, characteristics of the participants, content of the recordings, anonymization of the participants in the metadata, sharing conditions requested, to name a few. The different outcomes seem to be related to the possible interpretation of speech and voice samples as personal information, rather than to differences in the data and in our request form.

Whenever the speech signal is considered in itself personal information, the full anonymization of speech data is not feasible. It is also important to point out that the characteristics of child voice, that may allow the identification of an individual, change very quickly with age. This means that the participants will not be identifiable by their voice in just a few months or years after the recordings.

We hope that, in the future, uniform interpretations of regulation and practices in handling speech data will be introduced. We believe that the availability of open-access speech corpora with child speech (such as the proposed TeflonNorL2) is of essential value for scientific research and speech technology development, and they will bring about great advantages for the society.

8. Conclusions and Future Work

In this paper, we described the collection of Finnish, Norwegian and Swedish spoken by children L2 learners and children with SSD. We shared our experiences in recording, annotation and making the data publicly available. The data was collected for training ASR and APA systems to be used in mobile pronunciation learning games. The paper also contains details about the recorded speech data and the evaluation of the baseline ASR and APA systems. Due to the differences of the languages the results obtained by models trained for the various datasets are not directly comparable. We also make the data in the Norwegian TeflonNorL2 corpus publicly available. We believe that this will be a valuable resource for speech research and speech technology development.

The future work that has already started includes learning experiments in Finnish and Norwegian schools and Swedish speech therapy using the baseline ASR and APA models in the Pop2Talk

[5] https://live.european-language-grid.eu/ncc

[6] https://www.kielipankki.fi/language-bank/

language learning game. The game will also be used to collect more speech data to extend the existing datasets and improve the models. Although the game is currently only available to users for data collection, we plan to make Pop2Talk freely available, once the necessary computational resources have been secured long-term. We will also continue the efforts to make more of our speech data publicly available in near future. Furthermore, we are preparing a shared machine learning task and challenge in 2024 to develop better pronunciation rating algorithms for the Teflon-NorL2 data.

9. Acknowledgments

This research has been funded by the TEFLON NordForsk project nr. 103893.

10. Bibliographical References

A. Baevski, H. Zhou, A. Mohamed, and M. Auli. 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

J. De La Rosa, R.-A. Braaten, P. Kummervold, and F. Wetjen. 2023. Boosting Norwegian automatic speech recognition. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 555–564, Tórshavn, Faroe Islands. University of Tartu Library.

T. O. Engen and L. A. Kulbrandstad. 2004. *Tospråklighet, minoritetsspråk og minoritetsundervisning*. Gyldendal Akademisk.

Y. Getman, R. Al-Ghezi, T. Grosz, and M. Kurimo. 2023a. Multi-task wav2vec2 Serving as a Pronunciation Training System for Children. In *Proc. 9th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 36–40.

Y. Getman, N. Phan, R. Al-Ghezi, E. Voskoboinik, M. Singh, T. Grósz, M. Kurimo, G. Salvi, T. Svendsen, S. Strömbergsson, A. Smolander, and S. Ylinen. 2023b. Developing an ai-assisted low-resource spoken language learning app for children. *IEEE Access*, 11:86025–86037.

A. Hvenekilde. 1990. *Med to språk: Fem kontrastive språkstudier for lærere*. Cappelen.

R. Karhila, S. Ylinen, S. Enarvi, K. Palomäki, A. Nikulin, O. Rantula, V. Viitanen, K. Dhinakaran, A.-R. Smolander, H. Kallio, and M. Kurimo. 2017. SIAK a game for foreign language pronunciation learning. In *Proc. Interspeech*, pages 3429–3430.

O. Kommune. 2023. Tilbud til minoritetsspråklige elever (the rights to adapted language education for pupils from language minorities). Technical report, Oslo Kommune.

Opplaeringslova. 1998. Lov om grunnskolen og den vidaregåande opplæringa (lov-1998-07-17-61). Technical report, Lovdata.

L. Rumberg, C. Gebauer, H. Ehlert, M. Wallbaum, L. Bornholt, J. Ostermann, and U. Lüdtke. 2022. kidsTALC: A Corpus of 3- to 11-year-old German Children's Connected Natural Speech. In *Proc Interspeech*, pages 5160–5164.

J. Shi, N. Huo, and Q. Jin. 2020. Context-aware Goodness of Pronunciation for Computer-Assisted Pronunciation Training. In *Proc. Interspeech 2020*, Interspeech, pages 954–958, Shanghai. ISCA.

L. D. Shriberg and J. Kwiatkowski. 1982. Phonological disorders III: A procedure for assessing severity of involvement. *Journal of Speech and Hearing Disorders*, 47(3):256–270.

L. D. Shriberg, D. Austin, B. A. Lewis, J. L. McSweeny, and D. L. Wilson. 1997. The Percentage of Consonants Correct (PCC) Metric: Extensions and Reliability Data. *Journal of Speech, Language, and Hearing Research*, 40(4):708–722.

S. Strömbergsson, K. Holm, J. Edlund, T. Lagerberg, and A. McAllister. 2020. Audience response system-based evaluation of intelligibility of children's connected speech – validity, reliability and listener differences. *Journal of Communication Disorders*, 87:106037.

S. Sudhakara, M. K. Ramanathi, C. Yarra, and P. K. Ghosh. 2019. An Improved Goodness of Pronunciation (GoP) Measure for Pronunciation Evaluation with DNN-HMM System Considering HMM Transition Probabilities. In *Proc. Interspeech*, Interspeech 2019, pages 954–958, Graz. ISCA.

J. Vidal, L. Ferrer, and L. Brambilla. 2019. Epadb: A database for development of pronunciation assessment systems. In *Proc. Interspeech*, pages 589–593.

F. Yu, Z. Yao, X. Wang, K. An, L. Xie, Z. Ou, B. Liu, X. Li, and G. Miao. 2021. The slt 2021 children speech recognition challenge: Open datasets, rules and baselines. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 1117–1123.

J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang. 2021. speechocean762: An open-source non-native english speech corpus for pronunciation assessment. In *Proc. Interspeech*.

11. Language Resource References

C. Cucchiarini, H. V. Hamme, O. van Herwijnen and F. Smits. 2008. *Jasmin-spraakcorpus (Version 1.0)*. Available at the Dutch Language Institute: http://hdl.handle.net/10032/tm-a2-j7.

Khaldoun Shobaki, John-Paul Hosom and Ronald Cole. 1997. *CSLU: Kids‘ Speech Version 1.1, LDC2007S18*. Linguistic Data Consortium, ISLRN 965-489-670-052-2. [link].

Maxine Eskenazi, Jack Mostow and David Graff. 1997. *The CMU Kids Corpus LDC97S63*. Linguistic Data Consortium, ISLRN 566-795-587-797-8. [link].

Sameer Pradhan, Ronald Cole and Wayne Ward. 2021. *MyST Children’s Conversational Speech LDC2021S05*. Linguistic Data Consortium, ISLRN 848-818-101-134-5. [link].