9th Workshop on Speech and Language Technology in Education (SLaTE)
18-20 August 2023, Dublin, Ireland

New data, benchmark and baseline for L2 speaking assessment for low-resource languages

Mikko Kurimo[1], Yaroslav Getman[1], Ekaterina Voskoboinik[1], Ragheb Al-Ghezi[1], Heini Kallio[2], Mikko Kuronen[2], Anna von Zansen[3], Raili Hilden[3], Sirkku Kronholm[2], Ari Huhta[2] and Krister Linden[3]

[1]Department of Information and Communications Engineering, Aalto University, Finland
[2]University of Jyväskylä, Finland, [3]University of Helsinki, Finland
[1]{first.last}@aalto.fi, [3]{first.last}@helsinki.fi

Abstract

The development of large multilingual speech models provides the possibility to construct high-quality speech technology even for low-resource languages. In this paper, we present the speech data of L2 learners of Finnish and Finland Swedish that we have recently collected for training and evaluation of automatic speech recognition (ASR) and speaking assessment (ASA). It includes over 4000 recordings by over 300 students per language in short read-aloud and free-form tasks. The recordings have been manually transcribed and assessed for pronunciation, fluency, range, accuracy, task achievement, and a holistic proficiency level. We present also an ASR and ASA benchmarking setup we have constructed using this data and include results from our baseline systems built by fine-tuning a self-supervised multilingual model for the target language. In addition to benchmarking, our baseline system can be used by L2 students and teachers for online self-training and evaluation of oral proficiency.

Index Terms: ASR, L2 speaking assessment, wav2vec2.0, low-resource languages

1. Introduction

Automatic speaking assessment (ASA) supported by automatic speech recognition (ASR) is rapidly increasing its impact while the technology becomes more mature and effective and the applications more widespread. It can be used as a tool and additional resource in both self-learning, classroom teaching and education for human teachers and raters [1]. Its demand has recently increased even more by the geopolitical situation in Europe and the increased amount of immigrants.

The performance bottleneck in the machine learning-based ASA and ASR systems often comes from the shortage of suitable training data. While for many languages the available native speech data is abundant and even transcribed resources exist such as monologues [2], interviews, meetings [3] and parliament sessions [4], the publicly available L2 learners' speech data is much more limited. Moreover, collecting human expert assessments for the speakers' oral proficiency is expensive. What is more, every sample has to be rated by several experts, because they do not always agree.

Developing practical ASA systems for low-resource target languages such as Finnish and Swedish (more specifically, the Finland-Swedish variety) has not been possible, so far. There has not been enough transcribed training data for ASR development and rated speech data for ASA training. However, there have recently been many successful attempts to apply self-supervised deep acoustic transformer models like wav2vec2.0 [5] to low-resource domains including systems for ASR and various audio classification tasks [6, 7, 8, 9]. Inspired by the potential of the latest technology and the significance ASA may have for society, we have recently collected and annotated a significant amount of Finnish and Finland Swedish L2 learners' speech data in the DigiTala project. The target has been to cover as many skill levels as possible with as many speakers as possible. The transcription and rating effort has also been significant, although we have not obtained as wide coverage as we wished. Nevertheless, we have been able to create a reasonably large and useful data resource that will now be shared with the research community.

Typically, ASA systems are developed for L2 English where the test takers are abundant, the tests well established and resources high [10, 11, 12, 13]. However, their training data is rarely public, probably because of its commercial value or the privacy issues involved. For L2 English, a system based on pre-trained wav2vec2.0 was recently proposed in [14, 15].

Several open-access benchmarking data are available for tasks related to L2 speech, such as spoken language and accent recognition, accented speech recognition, spoken topic detection, and mispronunciation detection. However, according to the authors' knowledge, there is only one open dataset for L2 ASA, which is for the holistic proficiency level assessment of Asian Learners of English (ICNALE) [16]. Regarding Finnish and Finland Swedish, the National Certificates of Language Proficiency in Finland [17] records the test takers' speech, but even with their consent for research use, the data as such is not directly useful for training general ASA systems. This is because the test takers' speech is not transcribed, the oral proficiency is not rated separately for each task and the tasks vary based on the targeted proficiency level. Thus, we organized separate data collection campaigns in general upper secondary schools and universities in order to collect, transcribe and rate the speech data specifically for the ASR and ASA training.

The main contributions of our work include: 1. new carefully transcribed and rated L2 learners' speech data for two low-resource languages, 2. a setup for benchmarking ASR and ASA systems using the new data for training and evaluations, and 3. a state-of-the-art baseline system and its training and testing scripts as well as evaluation results for the benchmark. Parts of our data were used for evaluating our first wav2vec2.0 based systems in [18], and the baseline system and ASR error rates were already presented there. However, in this paper, we compare the ASA performance to inter-reviewer agreement, so all the tables and figures in this paper contain new results for the data and the baseline.

2. Data collection and annotation

The training and evaluation data for the L2 ASA systems were collected in general upper secondary schools and universities

10.21437/SLaTE.2023-32

with three main goals. The first goal was to reach as many Finnish and Finland Swedish L2 learners as possible to ensure the robustness of the final system for different voice characteristics and proficiency levels. For each language we managed to collect and assess over 4000 recordings by over 300 students as detailed in Table 1. Because no previous L2 speech data were available, the same collection provided also the training and evaluation for the ASR system on which the ASA was built.

	Fin		Swe	
	School	**Univ.**	**School**	**Univ.**
Duration, h				
Freeform	10.11	4.83	7.12	4.69
Readaloud	2.61	1.03	0	0.98
Total	12.72	5.86	7.12	5.67
Total (+unrated)	15.45	6.07	19.68	5.67
# of recordings				
Freeform	1336	1103	2025	1282
Readaloud	1186	780	0	959
Total	2522	1883	2025	2241
Total (+unrated)	3379	1965	4809	2242
# of tasks				
Freeform	20	10	22	11
Readaloud	6	7	0	8
Average Duration, s	18.15	11.20	12.67	9.11
# of ratings	5374	3986	4134	5223
# of students	202	113	181	120
# of raters	26	24	18	14
% of gender				
Female	54.4	41.6	n.a.	64.2
Male	41.9	56.6	n.a.	34.2
other/n.a.	3.8	1.8	n.a.	1.7
% of age group				
under 22	100	14.2	n.a.	28.3
22-26	0.0	48.7	n.a.	32.5
over 26	0.0	37.2	n.a.	39.2
% of L1				
Finnish	25.2		n.a.	
Swedish	27.4		n.a.	
Russian	10.6	8.4	n.a.	11.9
English		12.7	n.a.	10.4
German		7.6	n.a.	
Vietnamese		7.6	n.a.	13.3
not in Top3	36.7	63.6	n.a.	63.0

Table 1: *The data collection and rating results, including only the rated recordings. SweSchool has only Freeform speech rated so far. SweSchool also missed the speaker demographics, but we expect it to resemble FinSchool except almost all would have L1 Finnish. FinSchool also contains bilingual speakers who have L1 Finnish.*

The second goal was to include speech from a number of both read-aloud and free-form speaking tasks to capture different modes and fluency of speaking. The tasks were designed separately for different levels of L2 learners to simultaneously minimise the amount of data to annotate, but still cover the dimensions of speaking proficiency and be suitable for the participants. The number of tasks per dataset is listed in Table 1. The tasks are described in detail in [19], rating scales in [20]. For more details on data collection see [21, 22].

Because the human transcription and especially the assessment of L2 speech are laborious and expensive, we were able to provide only one human transcript and at least two human assessments for each recording, summing to over 9000 rated samples for each language (see Table 1). In addition to voluntary language teachers, the recruited raters were experts in language assessment, e.g. familiar with the National Certificates of Language Proficiency in Finland [17] or the Matriculation Examination. Each assessor participated in our rater training to improve the coherence of the ratings. To be able to analyze the coherence and quality of the ratings, the set of recordings selected for each assessor overlapped with the sets of at least two other assessors [21, 22].

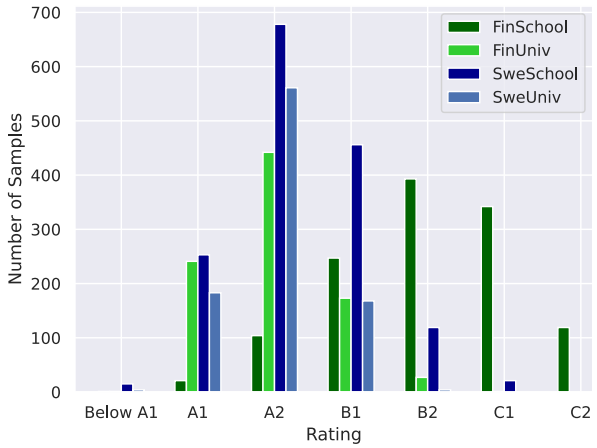

Figure 1: *The distribution of proficiency level ratings over the four different datasets.*

The third goal was not only to provide a CEFR-like holistic overall assessment (A1 – C2) for each recording, but also to take steps towards more explainable and transparent feedback to support self-learning. To be able to develop systems that could provide such further information, the same assessors rated the recordings separately on $0-3$ or $0-4$ scale on five different analytical dimensions: pronunciation, fluency, range, accuracy and task achievement. The rating criteria are described in detail in [20]. Figure 1 presents the distribution of the rated proficiency levels in the four different datasets that we collected: FinSchool, FinUniv, SweSchool, and SweUniv. The speaking tasks were tailored for each dataset to fit speakers of different amount of language studies. Interestingly, Figure 1 shows that in Swedish the rated proficiency levels overlap heavily compared to Finnish where the distributions are more distinct.

Table 1 and Figure 1 describe only the part of the collected data that was rated for oral proficiency in order to train and evaluate the ASA models and form the benchmark and baseline described in this paper. However, the speech data we will release in the Language Bank of Finland contains also other transcribed L2 speech material that we had collected and considered useful for training ASR models for L2 speakers. This other data contains the unrated part of the datasets mentioned in Table 1. This includes speech from additional read-aloud tasks of the same over 300 speakers per language, a total of 3 hours in Finnish and 12.5 hours in Swedish. It also contains data from the National Certificates of Language Profieciency (YKI) [23]: 9.5 hours in

Finnish (207 speakers) and 0.5 hours (24 speakers) in Swedish. The YKI data contains much longer recordings (about 100s) and the rating method is different than in our benchmark data.

3. The proposed benchmark and baseline

The primary goal of the DigiTala project was to develop ASR and ASA for L2 learners of the languages taught at the general upper secondary schools in Finland. The target was later extended to universities and from the assessment to self-learning tools. To make the data and tools openly accessible for research we have now collected the necessary consents and permits and the data will finally be available in the Language Bank of Finland `https://www.kielipankki.fi` in Autumn 2023. Unfortunately, the first dataset (SweSchool) was collected with consent from anonymous participants before GDPR was applied, so we can now only distribute the three newer datasets (FinSchool, FinUniv, SweUniv) described in Table 1. [1] The consent forms and background information templates are described in detail in [24].

To serve the primary goal of developing ASR and ASA with this data, we now suggest pre-defined benchmarks with training and evaluating sets in order to ensure the comparability of the resulting ASR and ASA systems. Because of the complexity and number of variables in the data, such as the tasks, levels, L2 learners and raters, and the limited size of the data, it was not possible to separate a single test that would be sufficiently balanced for all those variables. Thus, we decided to use four-fold cross-validation, where we have four sets with non-overlapping speakers that are balanced for tasks and proficiency. Each fold in turn is left out for testing, while the other three are utilized for training the models. In that way it is possible to aggregate the results of all four folds into the final evaluation set which is large enough for various analyses and to be used as a proper benchmark.

To make the proposed new benchmark more practical, we release also the ASR and ASA developed at Aalto University in 2023 and present their performance as the baseline for the benchmark. The models and training scripts will be shared also through our github `https://github.com/aalto-speech`. The main focus is on the Finnish ASR and ASA, because the Swedish ones utilized the part of our data (SweSchool) that can not be made openly available. However, in this paper, we provide also the Finland-Swedish baseline results for comparison and to show how the same method works on another language.

For the Finnish L2 ASR system, we selected the multilingual Fenno-Ugric wav2vec2-Large (317M parameters) [25] deep transformer model as the starting point. To adapt the speech model to Finnish, we first fine-tuned it with 100 hours from the colloquial Lahjoita Puhetta collection of spontaneous native speech [2]. Then, to adapt it to L2 learners' speech, we further fine-tuned it with Finnish DigiTala data (using the three folds selected for training) [18]. Unlike Finnish, Swedish has its own monolingual wav2vec2.0 model [26]. Because the preliminary experiments [6] indicated that the monolingual model work better for the target language than the multilingual one, we adopted it as our baseline. We then fine-tuned it directly with the SweSchool portion of the DigiTala data (the three folds selected for training) as in [18]. For the L2 ASA systems in both languages, we took the corresponding wav2vec2.0 systems fine-tuned for DigiTala ASR and trained the new classification heads to perform the ASA tasks as in [18].

[1] The unshareable part of the data was collected voluntarily for this project with insufficient documentation before the GDPR, i.e. the ASR and ASA models of the data can still be distributed as they contain no personal data.

	Fin	**Swe**
ASR		
WER, %	21.89	17.71
CER, %	7.06	9.08
ASA, holistic (CEFR)		
Range of classes	2-7	2-5
UAR, %	39.95	47.33
MAE		
human-to-human	0.782	0.613
machine-to-human	0.612	0.461
κ		
human-to-human	0.732	0.496
machine-to-human	0.807	0.524
ρ		
human-to-human	0.751	0.490
machine-to-human	0.803	0.524

Table 2: *Baseline results for ASR and ASA of the proficiency level. For Swedish this so far only covers the SweSchool dataset. Class range includes the levels with a sufficient amount of samples for evaluating the models. The metrics include word error rate (WER), character error rate (CER), uninterpolated average recall (UAR), mean absolute error (MAE), weighted quadratic Kappa (κ) and Spearman's correlation (ρ).*

The baseline ASR and ASA performance for the benchmark setups are presented in Table 2. The optimization of the character error rate (CER) is probably more relevant for the usefulness of ASR for pronunciation and fluency ratings, while the word error rate (WER) is more relevant for lexicon, grammar and task achievement. The reason for higher WER in Finnish, even though the CER is lower than in Swedish, is the more complex word composition of Finnish due to agglutination and richer morphology. To indicate the baseline performance in the ASA tasks, we present the uninterpolated average recall (UAR), mean absolute error (MAE), weighted quadratic kappa (κ) and Spearman's correlation (ρ). The lower recall and higher absolute error for Finnish is likely to be most affected by the wider range of classes available for the Finnish data compared to Swedish. We have also measured the inter-rater agreement for the benchmark setups as MAE, κ and ρ between the human ratings provided in the data.

Table 3 presents the performance in predicting the analytical dimensions of proficiency in addition to the holistic CEFR-like assessment. However, the results for the dimensions and languages are not directly comparable, because the narrower range and higher imbalance of classes may make some tasks appear easier than they are when compared to the others.

4. Discussion

The opportunity to develop effective ASR and ASA systems for low-resource tasks came from the recent advances in self-supervised training. Previously, it was inconceivable that there could be enough human transcribed and assessed data to train effective systems for Finnish and Finland Swedish. However,

ASA system		Class range	UAR, %	MAE	κ	ρ
Fluency						
Fin	HUM	2-4	-	0.575	0.393	0.392
	BL		55.67	0.359	0.507	0.522
Swe	HUM	1-3	-	0.425	0.498	0.490
	BL		59.41	0.305	0.560	0.574
Pronunciation						
Fin	HUM	2-4	-	0.445	0.513	0.531
	BL		54.62	0.269	0.583	0.612
Swe	HUM	2-3	-	0.419	0.162	0.162
	BL		67.53	0.343	0.276	0.290
Lexico-grammatical						
Fin	HUM	1-3	-	0.404	0.576	0.580
	BL		49.14	0.265	0.529	0.546
Swe	HUM	1-3	-	0.516	0.427	0.435
	BL		42.84	0.460	0.246	0.259
Task achievement						
Fin	HUM	1-3	-	0.410	0.340	0.298
	BL		44.09	0.318	0.365	0.390
Swe	HUM	1-3	-	0.621	0.376	0.371
	BL		59.51	0.320	0.650	0.714

Table 3: *Human-to-human (HUM) agreement and our baseline (BL) ASA results for all the assessed analytical dimensions. For Swedish this so far only covers the SweSchool dataset.*

our baseline results indicate that this is now possible by using the transcribed target data for fine-tuning the speech models trained on very large multilingual untranscribed native speech corpora. Thus, we expect that more accurate ASR and ASA systems can be developed in the near future when the pre-trained speech models become better. The technological advances will hopefully also encourage people to collect speech data for new low-resource tasks and new languages.

When using the pre-trained speech models as the starting point, the speech data in the benchmark seems to provide a sufficient ASR system to built an operational ASA system. Of course, the speech-to-text performance is still far from perfect and it may not be robust against noisy samples or new tasks. One known problem is that transcribing L2 learners' speech is a hard task even for humans, particularly for free-form speech including frequent hesitations. Another problem is that we can not use language models (LMs) in ASR, because they might also correct some of the errors produced by the speakers. However, with more computational power or in applications with less real-time constraints, we could make two ASR passes: the first one without the LM and the second one with it. The benefit would be a more readable transcript and a better starting point for the lexical, grammatical and task achievement assessment that would not be affected by the mispronunciations.

To improve the ASA system, the main limitation in our work is the available training data. Even if the human assessments were without errors, our data has severely imbalanced proficiency levels. This causes problems for training reliable classifiers, because the non-parametric models tend to learn the classes represented by more samples better. The imbalanced data is even more problematic for analysing the performance in specific speaking tasks or in the analytic dimensions (pronunciation, fluency, lexico-grammatical range and accuracy and task achievement) which are inherently more relevant to only a subset of the proficiency levels. One crucial limitation is also the accuracy of the human assessments for the short speaking tasks. While large-scale adding of more assessors per sample to eliminate the effect of human errors and outliers is expensive, we should maybe focus on how the human assessors are trained and how to make their task easier and less subjective by clarifying the tasks and rubrics.

From Table 3 it seems that the given baseline system already surpasses the inter-rater agreement in many criteria. However, the human-to-human and machine-human results are not fully comparable, because the automatic scores are compared to the average human scores whereas the human agreement is always between the two human raters. Thus, the more likely conclusion is that rating the short audio samples is a difficult task for human assessors, too.

The baseline ASA system presented in this paper is not very sophisticated in determining the correct proficiency level or analytical score when the human raters disagree. The correct proficiency level is needed both as the ASA training and evaluation target. Originally, we used the many-facet Rasch measurement [27] to compute fair averages from all the human ratings, but later we decided to just use a simple average rating, because in our preliminary experiments the effect on the final results was small, see also [22]. However, it would be interesting to use all our data now to study this more carefully and analyse the resulting differences, because the design of the overlapping ratings would make it possible.

The final limitation or problem with the data is also related to training the machine learning systems. The human assessors give their grades using a discrete scale of proficiency levels or score values that have been defined verbally [20], but averaging the grades or computing gradients produce decimal numbers which are not directly defined in the human scale. The approximative solution to map the classification task into a regression task is usually acceptable, but better solutions can be obtained by incorporating the inter-class distances into the machine learning algorithm [28].

5. Conclusions

In this work we present the L2 learners' Finnish and Finland Swedish speech data that we have collected, transcribed and rated in the DigiTala project. The data has been prepared for training and evaluation of low-resource L2 ASR and ASA systems. Together with the data, we will release a benchmark and our baseline system developed using four-fold cross-validation. In addition to performance comparisons, our baseline system can be used by L2 students and teachers as an online tool for self-training and evaluation of oral language proficiency. The tool is implemented in an ASR and ASA server that is accessed via a Moodle plugin [29] [2].

6. Acknowledgements

This work was done and the data were collected as part of the Academy of Finland grants number 322619, 322625, 322965 and 337073. The computational resources were provided by Aalto ScienceIT.

[2] https://github.com/aalto-speech/moodle-mod_digitala

7. References

[1] K. Evanini and K. Zechner, "Overview of automated speech scoring," in *Automated Speaking Assessment*. Routledge, 2020, pp. 3–20.

[2] A. Moisio, D. Porjazovski, A. Rouhe, Y. Getman, A. Virkkunen, R. AlGhezi, M. Lennes, T. Grósz, K. Lindén, and M. Kurimo, "Lahjoita puhetta: a large-scale corpus of spoken finnish with some benchmarks," *Language Resources and Evaluation*, 2022.

[3] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The AMI meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005, p. 100.

[4] A. Virkkunen, A. Rouhe, N. Phan, and M. Kurimo, "Finnish parliament ASR corpus: Analysis, benchmarks and statistics," *Language Resources and Evaluation*, 2023.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[6] R. Al-Ghezi, Y. Getman, A. Rouhe, R. Hildén, and M. Kurimo, "Self-Supervised End-to-End ASR for Low Resource L2 Swedish," in *Proc. Interspeech 2021*, 2021.

[7] X. Xu, Y. Kang, S. Cao, B. Lin, and L. Ma, "Explore wav2vec 2.0 for Mispronunciation Detection," in *Proc. Interspeech 2021*, 2021.

[8] L. Peng, K. Fu, B. Lin, D. Ke, and J. Zhan, "A Study on Fine-Tuning wav2vec2.0 Model for the Task of Mispronunciation Detection and Diagnosis," in *Proc. Interspeech 2021*, 2021.

[9] T. Grósz, D. Porjazovski, Y. Getman, S. Kadiri, and M. Kurimo, "Wav2vec2-based paralinguistic systems to recognise vocalised emotions and stuttering," in *Proceedings of the 30th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2022.

[10] X. Xi, D. Higgins, K. Zechner, and D. M. Williamson, "Automated scoring of spontaneous speech using SpeechRaterSM v1.0," *ETS Research Report Series*, vol. 2008, no. 2.

[11] Educational Testing Service. (2014) TOEFL iBT Speaking Section Scoring Guide.

[12] Pearson. (2017) PTE Academic Score Guide.

[13] J. Xu, E. Jones, V. Laxton, and E. Galaczi, "Assessing l2 english speaking using automated scoring technology: examining automarker reliability," *Assessment in Education: Principles, Policy & Practice*, vol. 28, no. 4, pp. 411–436, 2021.

[14] S. Bannò and M. Matassoni, "Proficiency assessment of l2 spoken english using wav2vec 2.0," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 1088–1095.

[15] S. Bannò, K. M. Knill, M. Matassoni, V. Raina, and M. J. F. Gales, "L2 proficiency assessment using self-supervised speech representations," 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2211.08849

[16] S. Ishikawa, "A new horizon in learner corpus studies: The aim of the icnale project," *Corpora and Language Technologies in Teaching, Learning and Research*, 01 2011.

[17] Finnish National Agency for Education. (2021) National Certificates of Language Proficiency. [Online]. Available: https://www.oph.fi/en/national-certificates-language-proficiency-yki

[18] R. Al-Ghezi, Y. Getman, E. Voskoboinik, M. Singh, and M. Kurimo, "Automatic rating of spontaneous speech for low-resource languages," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 339–345.

[19] A. von Zansen, "DigiTala's speaking tasks for L2 Finnish learners (proficiency level A, B1 and B2)," May 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6976044,https://doi.org/10.5281/zenodo.6562855,https://doi.org/10.5281/zenodo.6562865

[20] ——, "DigiTala's rating criteria: Holistic and analytic scales for assessing L2 speaking." 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6477089

[21] A. von Zansen, H. Kallio, M. Sneck, M. Kuronen, A. Huhta, and R. Hildén, "Ihmisarvioijien näkemyksiä suullisen kielitaidon automaattisesta arvioinnista, digitaalisesta arviointiprosessista sekä puhesuorituksista arvioitavista ulottuvuuksista: Human raters' perceptions of the automated assessment of oral language skills, the digital assessment process and the dimensions to be assessed from speaking performances," *AFinLAn vuosikirja*, pp. 370–394, 2022.

[22] A. von Zansen and A. Huhta, "Developing automated feedback on spoken performance: Exploring the functioning of five analytic rating scales using many-facet rasch measurement," in *Digital Research Data and Human Sciences*. Jyväskylän yliopisto, 2022.

[23] H. Kallio, S. Ohranen, T. Hirvelä, A. Huhta, A. von Zansen, Y. Getman, E. Voskoboinik, R. Al-Ghezi, M. Sneck, M. Kuronen, M. Kurimo, and R. Hildén, "DigiTala's YKI data." 2021. [Online]. Available: http://urn.fi/urn:nbn:fi:lb-2023012629

[24] A. von Zansen, "DigiTala's pre-test consent and background information form for L2 Finnish learners (upper secondary schools 2021)." May 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6562663

[25] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, vol. 1. Association for Computational Linguistics, Aug. 2021, pp. 993–1003.

[26] M. Malmsten, C. Haffenden, and L. Börjeson, "Hearing voices at the national library - a speech corpus and acoustic model for the swedish language," *CoRR*, vol. abs/2205.03026, 2022.

[27] J. Linacre, "Facets (version 3.83. 2)," 2020.

[28] E. Voskoboinik, Y. Getman, R. Al-Ghezi, M. Kurimo, and T. Grosz, "Automated assessment of task completion in spontaneous speech for finnish and finland swedish language learners," in *Proceedings of the 12th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2023)*, 2023.

[29] T. Alanen, J. Erkkilä, T. Harjunpää, and M. Heijala, "Digitala moodle plugin user manual," May 2022, The project's GitHub repository can be found and cited as: von Zansen, A., Alanen, T., Al-Ghezi, R., Erkkilä, J., Harjunpää, T., Heijala, M., Kallio, H. (2022). DigiTala Moodle plugin. https://github.com/aalto-speech/moodle- mod_digitala. [Online]. Available: https://doi.org/10.5281/zenodo.6535377