ASR-BASED AUTOMATIC ASSESSMENT OF ORAL PRODUCTION TASKS IN MULTILINGUAL CHILDREN

Eugenia Rykova[1,2], Tanja Rinker[1], Angela Grimm[3]

[1]Catholic University of Eichstätt-Ingolstadt, [2]University of Eastern Finland, [3]Goethe University Frankfurt
eugenia.rykova@ku.de

Abstract: The research project SPEAK aims to standardise a German language development test battery for multilingual children. This study evaluates the feasibility of automating a phonological production (nonword repetition, NWR) task and a vocabulary production task (Cross-linguistic Lexical Task, CLT) with the help of automatic speech recognition (ASR). The recommended accuracy threshold for ASR application in speech and language therapy is 80%. Five ASR models were tested with 858 audio recordings from the NWR task and 1267 audio recordings from the CLT noun production in German. Both tasks were administered to multilingual children and coded manually. The error rates (ERs) were calculated between the ASR transcriptions and the target, and the respective automatic phonemic transcriptions. For the CLT, the responses with the ER below 0.35 were accepted. For the NWR task, the transcriptions were post-processed to mimic the manual assessment process, and only the answers with ER = 0 were accepted. The mean accuracy of the automatic assessment was 69.7% for the NWR dataset, and 92.7% for the CLT dataset. High accuracy scores suggest the suitability of the ASR-based automatic assessment for evaluating multilingual children's performance in the noun-production task. The accuracy scores in the NWR task are still too low. Importantly for language diagnostics, the greatest drawback of automatic assessment (in both tasks) is misrecognising correctly produced items, which would not lead to underdiagnosis of children with language difficulties. This result would be in line with the proposals to accept overdiagnosis rather than underdiagnosis.

1 Background

In Germany and worldwide, the language abilities in multilingual children are often over- and underdiagnosed. On the one hand, monolingual norms generally do not reflect multilingual language development. On the other hand, a child might not have acquired enough language skills in the language in which they are diagnosed, usually societal and/or educational language (such as German in Germany) and there is no possibility to carry out diagnostics in their strongest language (home language), which becomes especially relevant in the context of high migration rates. Researchers propose to accept slight overdiagnosis (i.e., diagnosing typically developing children as disordered) rather than underdiagnosis (i.e., diagnosing language-disordered children as typically developing) given that underdiagnosis has more negative consequences for the children [1].

The research project SPEAK (*Sprachdiagnostik bei mehrsprachigen Kindern: Validierung einer Testbatterie* 'Language Diagnostics for Multilingual Children: Validation of a Test Battery') aims to standardise a language development test battery for multilingual children aged 4;0 to 8;11 (years;months) [2]. The test battery includes phonology, vocabulary, grammar, and narrative tasks. The final version will be available in the future for specialists under the name TEBIK 4–8 in the form of a serious game.

Phonology is tested with a quasi-universal nonword repetition (NWR task), which manipulates phonological complexity primarily through consonant clusters [3]. The task consists of a so-called (quasi)language-independent part and a so-called language-specific part.

Receptive vocabulary is assessed with a word-picture matching task, and productive vocabulary is assessed with a picture-naming task. Together, they form the Cross-linguistic Lexical Task (CLT) – in German, accordingly [4]. Possible difficulties with morphosyntax are identified with a German Sentence Repetition Task [5], and narrative and grammatical skills at a discourse level are evaluated with the Multilingual Assessment Instrument for Narratives [6].

Automation, at least in part, of the TEBIK 4–8 assessment will support the specialists and make the whole diagnostic process easier and quicker. All TEBIK 4–8 tasks, except receptive vocabulary, are based on oral speech production, which requires automatic speech recognition (ASR) for automation. McKechnie et al. [7] recommend an accuracy threshold of 80% for ASR use in speech and language therapy practice. Some applications for people with aphasia have surpassed this threshold in a picture-naming exercise – in other words, in a word verification task [see 8]. Following the methodological suggestions of Rykova and Walther [8], this study evaluates the feasibility of automating the phonological (NWR) task and the vocabulary-production (picture-naming) task from TEBIK 4–8 with off-the-shelf open-source ASR models.

2 Materials and Methods

2.1 NWR dataset

For the NWR task, 858 audio recordings served as the test material. They were collected from 13 multilingual children (9 girls) aged 5;6 to 6;6 years (mean, M = 6;0 years; standard deviation, SD = 0;4 years), who took part in the research project *cammino* [9]. The children had different language backgrounds in terms of spoken languages and age when they were first exposed to German. Two children were diagnosed with a developmental language disorder (DLD).

The participants repeated 66 pseudowords after hearing a pre-recorded audio target. Children's repetitions were recorded with a voice recorder; the recordings were semi-automatically segmented and manually labelled. The answers of the participants were also transcribed manually, and their performance was assessed by trained project members.

The language-independent items consist of typologically common vowels (/a/, /i/, /u/) and consonants (/p/, /k/, /f/, and /l/) combined into one- to three-syllable words, for example, /pli/ and /flipuka/. The language-dependent part introduces language-specific phonological structures, such as /s/ and /ʃ/ in onset and, to a lower extent, in coda positions, for example, /ʃpaklu/ and /kiʃ/. The idea behind the language-specific part is that, although /s/ + obstruent clusters are not exclusive to German, languages vary with regard to the representation of /s/ + obstruent clusters [10]. The item does not have to be produced fully identical to the target in order to be accepted as a correct answer. The following deviations are allowed:

1. voiced pairs of the target consonants (e.g., /b/ instead of /p/);
2. /s/ and /ʃ/ are interchangeable;
3. errors in the vowel production do not count towards the final accuracy.

2.2 CLT dataset

For the vocabulary-production task, a total of 1308 audio recordings served as the test material. These data were collected from 41 multilingual children (29 girls) aged 4;0 to 8;9 years (M = 6;3 years, SD = 1;2 years), who took part in a research project assessing language development in pre-school and school children [11]. The children had different language backgrounds. No comprehensive information on possible DLD diagnoses was available.

The participants performed the German CLT with an iPad, and their answers were recorded within an app, separately for each item. The test administrator also kept a paper protocol, in which they recorded children's answers and their correctness. In this feasibility study, only the noun-production task is included because there is more variability in verb responses, and its evaluation needs a more careful procedure.

The app automatically records only 20 seconds after the picture stimulus is presented. However, a participant might give a correct answer after the recording stops, which would be marked as correct in a paper protocol, but not reflected in the audio recording. To avoid these mismatches, those audio recordings that had been erroneously evaluated with an ASR-based method were inspected auditorily, and 41 of them were deleted. The final dataset comprises 1267 audio recordings.

The noun-production task consists of 32 items from different semantic categories, mostly from such categories as food (*Banane* 'banana', *Zwiebel* 'onion'), animals (*Eule* 'owl', *Schnecke* 'snail'), and everyday objects (*Socke* 'sock', *Rutsche* 'slide'). The length of the words ranges from one to three syllables. Compound hyponyms are accepted as correct answers when they correspond to the picture (e.g., *Haarbürste* 'hairbrush' for target *Bürste* 'brush'). Otherwise, two-thirds of the item should be produced correctly and not correspond to a semantically different word (e.g., *Tür* 'door' is not accepted as a correct answer for target *Tor* 'gate').

2.3 Automatic transcriptions

Four ASR models that have previously shown the best results in single-word recognition of atypical speech [8] were used in this study. Three of them are fine-tuned with German datasets versions of Facebook Wav2Vec2 models, and one is a large version of the Conformer-Transducer model trained on German data. Additionally, a small (default) Whisper model [12] was tested, with the output language set to German.

Each audio recording was transcribed with every model, and for each ASR transcription an automatic grapheme-to-phoneme (g2p) transcription was generated with CoML's eSpeak-based phonemizer [13]. Following the suggestions from [8], besides raw ASR output, each separate word (a sequence from space to space) from the output and a sequence of all the words without spaces between them were analysed. For each of the analysed units a g2p transcription was generated.

2.4 Transcription evaluation

The character error rate (CER) was calculated between the ASR output (each possible option) and the target word, and the phonemic error rate (PER) was calculated between the respective g2p transcriptions. Both error rates (ERs) were normalised.

For the NWR task, the transcriptions were post-processed to unify pairs of voiced and unvoiced consonants, and alveolar and postalveolar fricatives; and eliminate the distinction among the vowels mimicking the manual assessment process. For example, an orthographic form <klifapu> was converted to KlAFAPA, where K stands for <k> or <g>, F stands for <f>, <v> or <w>, P stands for <p> or <b>, and A stands for any vowel. The ERs were calculated after post-processing and only the answers with ER = 0 were accepted.

For the CLT, the minimum of the ERs across the transcriptions was selected and compared to an ER threshold. The threshold was set to 0.35 upon discussion with the CLT creators – in other words, at least 65% of the word must be pronounced correctly (or recognised by the ASR) in order for the answer to be considered correct, which corresponds to the rules of manual assessment.

3 Results

3.1 NWR task

The accuracy of the automatic assessment ranged from 37.9% to 97% among the participants, with a mean value of 69.7%, and a standard deviation of 14%. The results per participant are presented in Figure 1. The children are ordered according to their performance in the NWR

task, and this score is given at the bottom of each bar. The speakers are labelled according to their gender (*f* for girls, *m* for boys) and age (in months), and those diagnosed with DLD are marked with an asterisk.

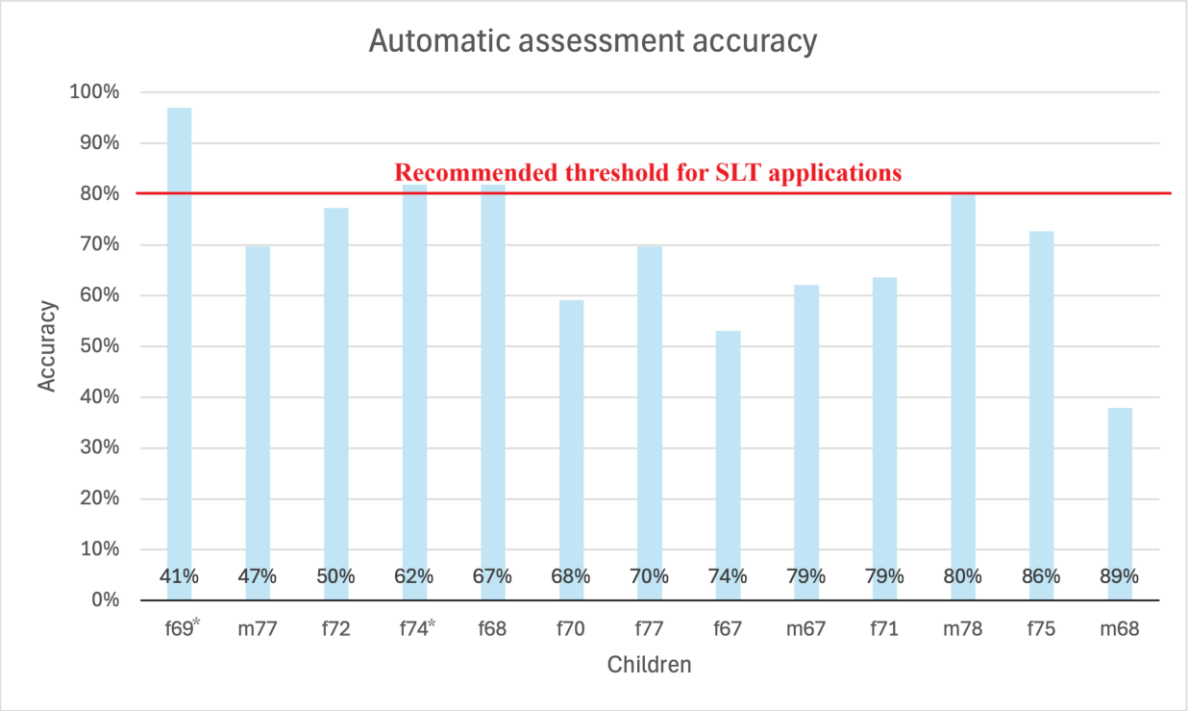

Figure 1 – Automatic assessment accuracy for the NWR-task dataset

No significant correlation was found between either the children's age or their length of exposure to German and the accuracy of the automatic assessment. A moderate negative correlation was observed between children's performance in the NWR-task and the accuracy of the ASR-based assessment: $r(11) = -0.63$, $p = 0.022$.

3.2 CLT

For the CLT dataset, the accuracy of the automatic assessment ranged from 75% to 100% among the participants, with a mean value of 92.7%, and a standard deviation of 6.3%. The results per participant are presented in Figure 2. The children are ordered according to their performance in the test, and this score is given at the bottom of each bar. The speakers are labelled according to their gender (*f* for girls, *m* for boys) and age (in months). No significant correlations were found between either the children's age or their performance in the test and the accuracy of the automatic assessment.

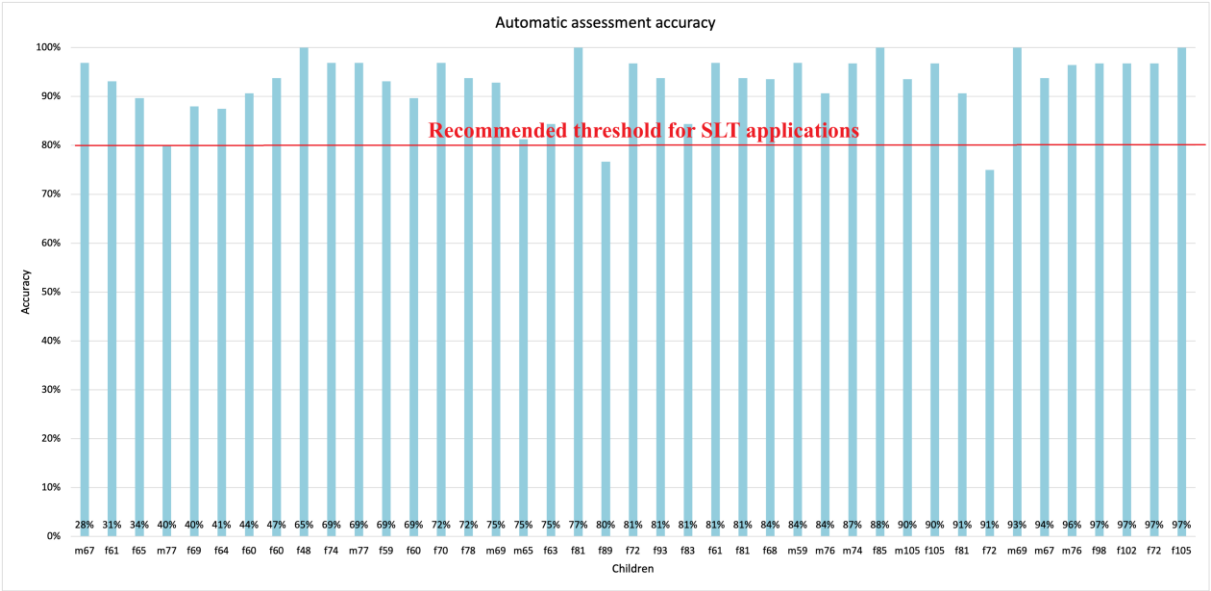

Figure 2 – Automatic assessment accuracy for the CLT dataset

3.3 General observations

A more detailed analysis of ASR-based automatic assessment is presented in Figure 3 in the form of confusion matrices for each dataset.

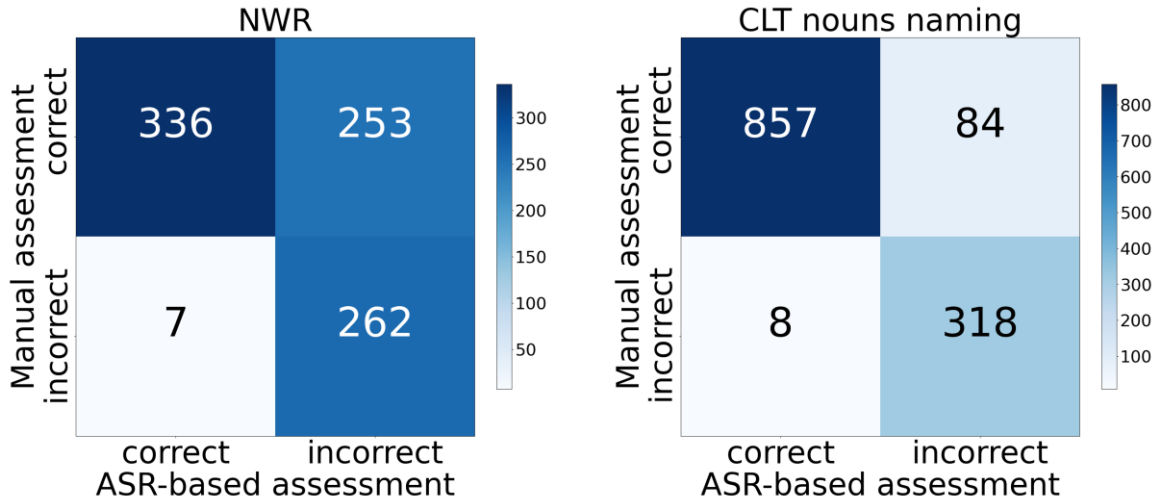

Figure 3 – Details on ASR-based assessment

In both tasks, the precision of recognising correct items and the recall of incorrect items are above 97%. The recall of correct items is 57% for the NWR task and 91% for the CLT; and the precision of recognising incorrect items is 50% and 79%, respectively.

While in the NWR task both false positives and false negatives result from ASR imperfections (e.g., mixing unvoiced plosives with each other or missing some sounds), false positives in the CLT have a different nature and result from orthographic and/or phonemic proximity of the target and erroneous answer (e.g., answer *Tür* 'door' for target *Tor* 'gate', which accounts for five out of eight false positives). Adding a word-check component would help reduce such cases [cf. 8].

A quick look at the ASR models' performance revealed that the Whisper model was the only one that correctly assessed 3.1% of the NWR data, and the one with the selected minimum score for the 4.3% of the CLT data.

4 Discussion

High accuracy scores suggest the suitability of the ASR-based automatic assessment for evaluating multilingual children's performance in the TEBIK vocabulary (noun)-production subtest. In the phonological subtest, a nonword repetition task, the accuracy scores are still too low. This task is more challenging for ASR because the models cannot use lexical information in recognition of nonwords. Furthermore, the threshold for evaluation is set to 0 (cf. 0.35 for the CLT transcription evaluation), which does not allow any inaccuracies from the ASR. Setting a higher threshold would lead to a greater number of false positives and possible underdiagnosis of language difficulties.

Importantly for language diagnostics, automatic evaluation is conservative with respect to false negatives in both tasks, and in the NWR task there is a slight tendency toward better automatic assessment of low performers. In other words, speakers' errors are not overlooked, so that the possible presence of language disorders would not be underdiagnosed but rather overdiagnosed – in line with the recommendations for language diagnostics [1]. However, such a high possibility of overdiagnosis as observed in the presented automatic assessment of the NWR task is not desirable and is therefore subject to further improvements.

Item-based analysis of automatic assessment errors could reveal the nonwords particularly challenging for ASR. In the TEBIK phonological subtest, the number of test items will be reduced, which could lead to elimination of such challenging items and make automatic assessment more reliable. For further improvements, adapting the ASR models is foreseen.

Despite being unsuccessful with speech data collected from speakers with aphasia [8], the Whisper model has been a valuable addition to the tested children's speech datasets. Large(r) Whisper models should be included in future experiments with the TEBIK data, together with speech-to-IPA (International Phonetic Alphabet) models.

At the level of vocabulary, evaluation of the automatic assessment of verb production is seen as the next step. Next, testing the suitability of ASR-based assessment – possibly in combination with natural language processing methods – will be performed for the other TEBIK subtests. Finally, integration of the speech and language technologies into the serious game for the online assessment should be carried out.

5 References

[1] WINTER, K.: *Numbers of bilingual children in speech and language therapy*. In *International Journal of Bilingualism*, 5(4), pp. 465 – 495, 2001.

[2] GAGARINA, N., A. GRIMM, T. RINKER, and A.-L. SCHERGER: *Testbatterie zur Diagnostik von Sprachentwicklungsstörungen bei mehrsprachigen Kindern in Deutschland [Test battery for the diagnosis of developmental language disorders in multilingual children in Germany]*. In *LOGOS*, 33 (3), pp. 186 – 192, 2025.

[3] GRIMM, A.: *The use of the LITMUS quasi-universal nonword repetition task to identify DLD in monolingual and early second language learners aged 8 to 10*. In *Languages*, 7(3), p. 218, 2022.

[4] RINKER, T. and N. GAGARINA: *CLT – Crosslinguistic Lexical Task – Deutsche Version [German Version]*. Universität Konstanz & ZAS Berlin, 2014.

[5] HAMANN, C., S. CHILLA, E. RUIGENDIJK, and L. ABED IBRAHIM: (2013). *A German sentence repetition task: testing bilingual Russian-German children,* poster presented at the COST Action IS0806 Conference, Krakow, 2013.

[6] GAGARINA, N., D. KLOP, S. KUNNARI, K. TANTELE, T. VÄLIMAA, U. BOHNACKER, and J. WALTERS: *MAIN: Multilingual Assessment Instrument for Narratives – Revised*. In *ZAS Papers in Linguistics*, 63, 20, 2019.

[7] MCKECHNIE, J., B. AHMED, R. GUTIERREZ-OSUNA, P. MONROE, P. MCCABE, and K. J. BALLARD: *Automated speech analysis tools for children's speech production: a*

systematic literature review. In *International Journal of Speech-Language Pathology*, 20(6), pp. 583 – 598, 2018.

[8] RYKOVA, E. and M. WALTHER: *Evaluation of German ASR solutions in the context of speech and language therapy support of people with aphasia*. In *Loquens*, 12, e116, 2025.

[9] SCHULZ, P., A. GRIMM, B. GEIST, and B. VOET CORNELLI*: cammino – Mehrsprachigkeit am Übergang zwischen Kita und Grundschule [Multilingualism at the transition between preschool and primary school].* In BUNDESMINISTERIUM FÜR BILDUNG UND FORSCHUNG (ed.): *Bildungsforschung 2020 – Herausforderungen und Perspektiven*, pp. 281 – 284. Berlin: BMBF, 2014.

[10] FIKKERT, P., and M. J. FREITAS: *The role of language-specific phonotactics in the acquisition of onset clusters.* In *Linguistics in the Netherlands*, 21, pp. 58 – 68, 2004.

[11] BAUMGARTNER, R. and T. RINKER: *Mehrsprachigkeit am Übergang Kita – Grundschule [Multilingualism in the transition from kindergarten to primary school].* In: U. SCHRÄPLER & A. BLECHSCHMIDT (eds.): *Mehrsprachigkeit in Logopädie und Unterricht*, pp. 127 – 143. Schwabe, Basel, 2018.

[12] RADFORD, A., J.W. KIM, T. XU, G. BROCKMAN, C. MCLEAVEY, and I. SUTSKEVER: *Robust speech recognition via large-scale weak supervision.* In *Proceedings of the 40th International Conference on Machine Learning*, 202, pp. 28492 – 28518, PMLR, 2023.

[13] BERNARD, M. and H. TITEUX: *Phonemizer: text to phones transcription for multiple languages in Python*. In *Journal of Open Source Software*, 6(68), 3958, 2021.