Towards a Reliable French Speech Recognition Tool for an Automated Diagnosis of Learning Disabilities

Jihene Rezgui[1,2], Félix Jobin[1], Younes Kechout[1,2], Chritine Turgeon[3] and Foutse Khomh[2]
[1] **L**aboratoire **R**echerche **I**nformatique **M**aisonneuve (LRIMa) Montreal, Canada,
[2] Polytechnique Montreal, Canada, [3] Université LAVAL, Québec, Canada
jrezgui@cmaisonneuve.qc.ca, foutse.khomh@polymtl.ca, christine.turgeon@fmed.ulaval.ca

Abstract –Dyslexia, characterized by severe challenges in reading and spelling acquisition, presents a substantial barrier to proficient literacy, resulting in significantly reduced reading speed (2 to 3 times slower) and diminished text comprehension. With a prevalence ranging from 5% to 10% in the population, early intervention by speech and language pathologists (SLPs) can mitigate dyslexia's effects, but the diagnosis bottleneck impedes timely support. To address this, we propose leveraging machine learning tools to expedite the diagnosis process, focusing on automating phonetic transcription, a critical step in dyslexia assessment. We investigated the practicality of two model configurations utilizing Google's speech-to-text API with children speech in evaluation scenarios and compared their results against transcriptions crafted by experts. The first configuration focuses on Google API's speech-to-text while the second integrates Phonemizer, a text-to-phonemes tool based on a dictionary. Results analysis indicate that our Google-Phonemizer model yields reading accuracies comparable to those computed from human-made transcriptions, offering promise for clinical application. These findings underscore the potential of AI-driven solutions to enhance dyslexia diagnosis efficiency, paving the way for improved accessibility to vital SLP services.

Keywords: automatic speech recognition, reliability, phoneme recognition, learning disabilities, dyslexia.

I. INTRODUCTION

According to the World Federation of Neurology, dyslexia is an "impairment of reading learning occurring in spite of normal intelligence, absence of sensory disorders, adequate schooling, adequate socio-cultural opportunities". It is a neurological and strong hereditary disorder favored by certain environments. This Specific Learning Disability (SLD), impacting around 10% of the total population, creates unfair challenges in every aspect of life since reading is a required skill for nearly all activities in the modern world [1-4]. Dyslexia and other learning disabilities have also been reported to have correlations with other psychological issues, such as emotional/behavioural problems and Attention Deficit Hyperactivity Disorder (ADHD), which induces more challenges in everyday life [5].

Side-effects of dyslexia can be reduced when children are taken in charge by Speech and Language Pathologists (SLPs) or other rehabilitation professionals. When they become involved early in the child's development, SLPs can suggest targeted exercises to improve child's learning skills [6]. However, this constant follow-up is impossible without a first diagnosis from one of these professionals. Since SLPs are usually overwhelmed by the population's demand for learning disabilities diagnosis [7], it makes this step a choking point in the rehabilitation process of children living with dyslexia. Therefore, it is crucial to improve accessibility to SLP services by accelerating the diagnosis process first.

In current practice, SLPs lack tools to enhance the speed and accuracy of core tasks in diagnosing dyslexia, with one of the most challenging tasks being the phonetic transcription of speech. In a standard evaluation, the SLP asks his client to read out loud specific texts and records it. He then has to manually identify every phoneme produced by the client to produce the phonetic transcription, which then can be compared to an expected transcription to find anomalies in the client's speech. To the best of our knowledge, despite advanced research in phoneme recognition, **there is no technology specifically applied to ease the phonetic transcription process that is mandatory for every dyslexia diagnosis** [8]. Artificial intelligence (AI) emerges as a promising avenue to bridge the notable disparity in existing practices regarding the phonetic transcription process in dyslexia diagnosis [9]. In this paper, we delve into potential solutions leveraging off-the-shelf AI models to get a sense of what is possible to achieve with available solutions, such as Google API.

Our contributions in this paper can be summarized as follows: (1) We gathered French audio samples from children in the perspective of Automatic Speech Recognition (ASR) with children voices thanks to PhonIA tools [10]; (2) we labeled them with a phonetic transcription with the help of language specialists; (3) we explored solutions to automate a core step of the dyslexia diagnosis; (4) we integrated Google's speech-to-text model with a phonemizer and compared it with its traditional API and (5) assessed the use of ASR solutions in a clinical setting of dyslexia diagnosis.

Section II gives a brief overview of related works in automated SLD detection with speech and other inputs. Section III describes transcription models and their different components. Section IV explains the dataset used in this work and the data collection steps. Section V shows the ML Models performances as well based on

Google's model under several criteria. Finally, Section VI concludes the paper and provides potential paths for future work.

II. RELATED WORK

Most of the related work in automated SLD detection [11-23] can be placed in the flowchart of Fig. 1. Most popular angles to tackle this problem include end-to-end solutions [11-12] to automate the entire process with machine learning (ML). Other researchers have also explored models based on eye tracking and handwriting to achieve the same objective [13-23].

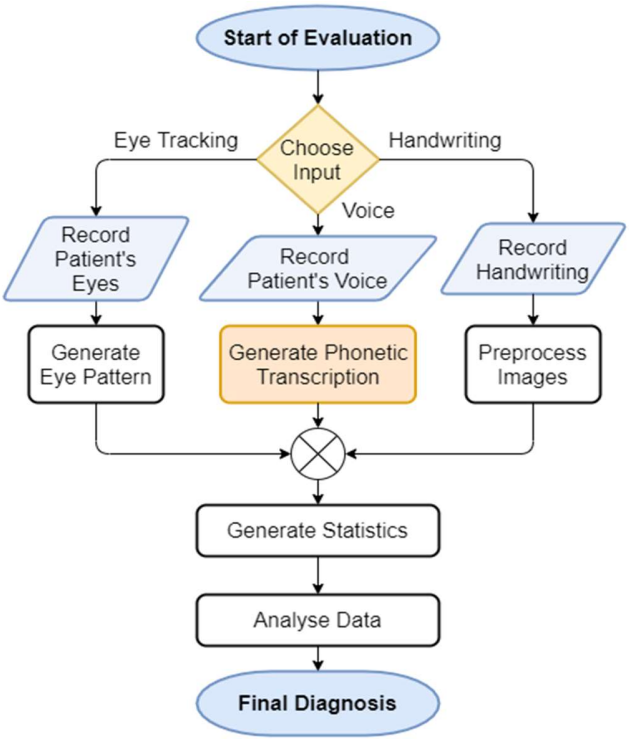

Fig.1. Flowchart of the evaluation process. The orange step (phonetic transcription) is the main focus of this work.

A. End-to-End Solutions

End-to-end solutions make the hypothesis that a ML model could learn to find patterns in the patient's speech that would be related to a specific learning or speech disability. These ASR systems take raw speech or speech features as inputs and generate a prediction from a range of classes where each of them represents a severity level of the potential disability [11]. In other words, all-in-one systems take the patient's voice as input and jump directly to a preliminary diagnosis. Some of these works have shown impressive results, reaching more than 95 % accuracy for specific models [12].

End-to-end solutions are promising for broad screening applications. For example, we could use such models to assess if a child needs to go through a whole evaluation process with an SLP. However, they could not be used to accelerate the evaluation process, because systems delivering decisions such as an SLD diagnosis must be highly explainable and understandable by the professional to be trusted. To ensure that the diagnosis is as reliable as possible, it is best to focus on steps that can be automated within the process in order to leave the final decision to the professional.

B. Solutions with Other Inputs

Literature has shown that some eye patterns can be associated with dyslexia and other SLD [13-14]. With that in mind researchers have conceived ML systems to predict these kinds of disabilities based on eye tracking records in various contexts [15]. For example, [16] have achieved an accuracy of 86.25% with a support vector machine model with input data coming from eye tracking of adolescents watching paintings. Some researchers have explored images of handwriting as input data since it has been shown to be correlated with dyslexia [18]. With this approach, [19] have reported an accuracy of 96.4% when compared to the true SLD diagnosis of the child. Finally, another approach has been explored by researchers and involves the use of electroencephalography (EEG) signals as inputs [17]. In summary, these approaches also propose solutions covering from the input record to the final diagnosis, but they use other relevant inputs.

In this work, we wanted to explore ML tools that would be useful for SLD professionals while being as close as possible to their current practice. Since the phonetic transcription generation is the most popular approach and is currently used in clinics to diagnose SLDs in Quebec [20], we decided to prioritize this approach.

III. TRANSCRIPTION MODELS

As base components of tested solutions, we selected off-the-shelf models to measure their potential efficiency in the work of SLPs. Results will serve as references to compare other solutions that will focus on the same objective. In this section, we introduce and compare two transcription models utilized in our study.

A. Google API Model

The primary objective of the Google API Model is to provide a straightforward and accessible form of ASR, which can potentially serve the needs of SLPs in diagnosing dyslexia. We selected this model to get a sense of the performance achieved by a broadly available model. The Google speech-to-text API is powered by Chirp, a version of Google's Universal Speech Model (USM) which is a 2-billion parameters speech model trained on millions of hours of audio in more than 100 languages [21].

This model initiates the transcription process with raw audio data obtained from reading one of the four tasks. Subsequently, the Google API is employed to generate text transcriptions from these audio inputs. The accuracy of the child's reading is then evaluated by comparing the generated transcription to the original text. The complete flowchart of this model can be compared to the other one in Fig. 2.

B. Google-Phonemizer Model

Our proposed Google-Phonemizer Model shares the same objective as the Google API Model but incorporates an additional layer of complexity by integrating the Phonemizer into the transcription process. The phonemizer utilized in this model is a comprehensive Python library translating text into phonemes of a specific language and region [22]. The phonetic transcription is based on a range of dictionaries, from which we chose eSpeak for its versatility across multiple languages [23]. Since Canadian French was not accessible with this dictionary, we selected France's accent. When generating expected phonetic transcriptions from each text to be read by children, we validated all transcriptions with a professional SLP that judged their exactness based on his experience. This evaluation confirmed the usability of the generated transcriptions for Canadian French.

After generating text transcriptions using the Google API, the phonemizer is applied to produce phonetic transcriptions from the text. The accuracy of the transcription is then assessed by comparing the generated phonetic transcription to the expected phonetic transcription derived from the read text and validated by the SLP. As with the Google API Model, differences between the expected and generated transcriptions are utilized to gauge the accuracy of the child's reading. The model's flowchart can be compared with the previous model in Fig. 2.

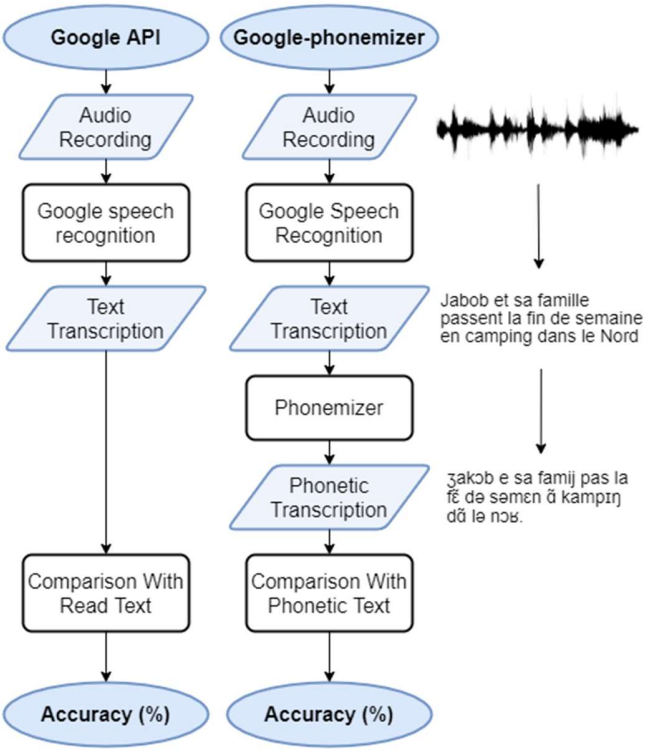

Fig.2. Flowchart of tested models with French data. Each model generates the text transcription of the audio, but Google-phonemizer also generates the phonetic transcription based on the text. Both models output an accuracy score when compared with the expected transcription.

C. Accuracy Significance and Calculation

SLPs use the phonetic transcription of speech to compute statistics related to the overall child's reading fluency, one of them being the number of pronunciation errors. These errors can be split into 2 categories: (a) added words (AW) and (b) removed words (RW). AW are words added in the speaker's transcription, while RW words removed from the original transcription. Words can be considered added or removed when the same words (e.g. the same series of characters between two spaces) cannot be found in the compared text at a similar position relatively to previous and following text. This metric definition allows us to compare models between them even if their output is not exactly of the same nature.

Accuracy in reading can be defined as the number of correctly read words over the total number of words read. As shown in Eq. (1), we obtain the number of read words (W_{read}) by subtracting the number of RW from the total number of words (W) in the transcription. After that, we subtract W_{read} with AW, as shown in Eq. (2). The result is then put into relation with W_{read} to get a percentage of accuracy.

$$W_{read} = W - RW \qquad \text{Eq.(1)}$$

$$Acc\ (\%) = \frac{W_{read} - AW}{W_{read}} \times 100 \qquad \text{Eq.(2)}$$

In clinical reports, accuracy values are compared with standardized values, which are determined by gathering results from children of the same age and scholarship level for the specific test. SLPs refer to those values to assess if the result is expected from a child in its state of development. In this work, we focus on the accuracy result because the conclusions are expected to be the same for an identical accuracy. Fig. 3 displays the relation between all metrics of interest to compare transcription models.

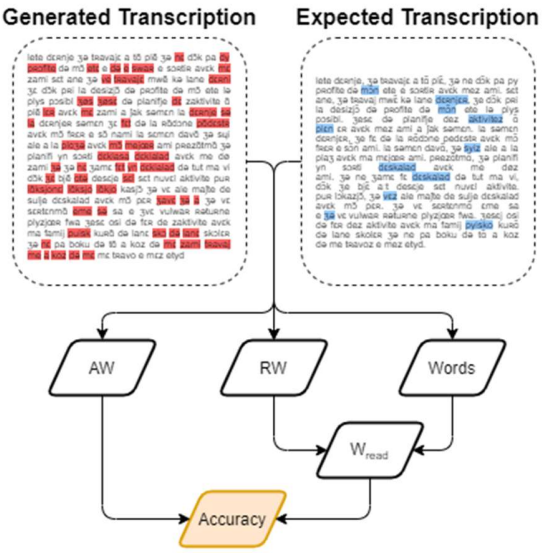

Fig.3. Relation between all metrics of interest in this work. Red highlights are AW in the generated transcription while blue highlights are RW in the generated transcription.

IV. DATA COLLECTION

To assess the reliability of selected models, we compared output transcriptions with a human-made phonetic transcription. Unfortunately, French audio datasets including phonetic transcriptions are almost non-existent publicly, which encouraged us to work with a young company specialized in the field.

PhonIA [10] is a startup whose mission is to accelerate the evaluation and follow-up process of SLPs with innovative technologies such as ASR. For this purpose, we utilized their tool to gather an in-house dataset of child speech while executing a reading task. Recordings were collected using one of the two following methods: in-person with the presence of a linguistics master student, or in an uncontrolled remote environment. This inconsistency was chosen to test the reliability and robustness of tested models. We ensured that selected recordings were still understandable to be able to produce a phonetic transcription out of them.

In this work, we used 167.56 minutes of audio recording, which corresponds to 119 samples from 40 French Canadian children ranging from 7 to 12 years old. Four different French texts written by a practicing SLP were used as reading tasks and each sample uses one of them. Table 1 details information about each reading task. For example, the text of the task "Sentences 1" (See Table.1) in French, which has been read by children in some samples, is shown in Fig.4.

Table 1. Reading task information. Expected phonemes were set by transcribing with Phonemizer and were validated by a SLP.

Name	Type	Number of Sentences	Number of Words	Expected phonemes
Text 1	Text reading	12	189	658
Sentences 1	Sentences reading	14	152	554
Text 2	Text reading	10	163	600
Sentences 2	Sentences reading	12	133	448

All samples used in this work have their corresponding phonetic transcription written by 2 language stimulation agents. They have an academic background in linguistics and are trained to produce phonetic transcriptions in French. Even if the read text was the same for many samples, we ensured that each transcription was only based on the audio signal. These precautions were taken because we wanted to reduce the bias a transcriber could have by starting a transcription with another one that is based on the same task, but not the same voice. To help SLPs and other professionals in the SLD evaluation process, ASR systems will need to adapt to any accent they could encounter in a specific language. This procedure is key to ensure that the speaker's accent is reflected in the transcription, resulting in a better representation of the speech.

Je vais nager dans la piscine de mon voisin Tom mercredi prochain.
Je bois de l'eau quatre fois par jour afin d'être bien hydraté.
Mon passe-temps préféré est le chant.
Mon cousin a animé une émission de radio la semaine passée.
Je vais magasiner une robe de bal avec ma grand-maman.
Pour Noël, j'aimerais recevoir une raquette de tennis orange.
J'ai perdu ma bouteille d'eau à l'école de ma petite sœur.
Mon oncle se marie le mois prochain et je n'ai pas de souliers à porter.
Mon ordinateur redémarre souvent, donc j'aimerais le changer.
Les yeux de ce bébé sont ronds et bruns.
La robe de la petite fille devant moi est rose et fleurie.
L'auto rouge de mon voisin est au garage depuis trois jours.
Mon rendez-vous chez le dentiste était très douloureux aujourd'hui.
Le gardien de mon équipe de soccer était malade hier donc nous avons dû trouver un remplacement.

ENGLISH TRANSLATION:

I'm going swimming in my neighbor Tom's pool next Wednesday.
I drink water four times a day to stay well hydrated.
My favorite hobby is singing.
My cousin hosted a radio show last week.
I'm going shopping for a prom dress with my grandmother.
For Christmas, I would like to receive an orange tennis racket.
I lost my water bottle at my little sister's school.
My uncle is getting married next month and I don't have any shoes to wear.
My computer restarts often, so I would like to change it.
This baby's eyes are round and brown.
The dress of the little girl in front of me is pink and flowery.
My neighbor's red car has been in the garage for three days.
My dentist appointment was very painful today.
My soccer team's goalie was sick yesterday so we had to find a replacement.

Fig.4. Text of the task "Sentences 1" in French, which has been read by children in some samples. The English translation is written after it.

In terms of data pre-processing, we needed to ensure that characters used by Phonemizer were the same as the ones used by speech specialists to avoid unwanted differences. As an example, we found that two different characters (Unicode codepoints U+025B and U+03B5) were used to identity [ɛ], which were considered different by the algorithm while having the same shape. Punctuations marks and capital letters were also removed to avoid identifying differences with points, commas, and apostrophes.

V. RESULTS ANALYSIS

To assess the performance of each model at finding the reading accuracy, we compared each accuracy with the accuracy calculated using the human-made transcription. We then computed the deviation percentage where we considered the human-made transcription as the reference. Average deviations have been computed by taking the deviation percentage of accuracies from each sample with respect to their human-made counterpart and by taking the average of all deviations, as shown in Eq.(3). N is the total number of samples while $Acc_{g,i}$ is the generated accuracy of the ith sample and $Acc_{e,i}$ is the expected accuracy of the same sample (i.e. calculated with the human-made

transcription). This metric is used to better understand the general disparity of transcriptions between ASR and human-made versions.

$$Dev\ (\%) = \frac{1}{n}\sum_{i=1}^{n}\frac{|Acc_{g,i} - Acc_{e,i}|}{Acc_{e,i}} \times 100 \qquad \text{Eq.(3)}$$

We also considered the number of samples with dominant added words (DAW) and dominant removed words (DRW). DAW corresponds to the number of transcriptions with AW as the prominent source of differences while DRW is the same for RW. The addition of these statistics allowed us to better understand each model's behaviour in terms of transcription generation, as whether a model is more prone to add or remove segments. Table 2 compares both models with human-made transcription results.

What first comes to light in the Table.2 is the greater proximity of accuracies from Google-Phonemizer. When compared to human-made accuracies, Google API can reach a deviation of 20.47% while Google-phonemizer is able to achieve 8.68% of deviation. Considering that both solutions were made from off-the-shelf models, the ability to reach this kind of closeness from the second model shows that there is a strong potential in the use of these tools in phonetic transcriptions. It is also important to note that human-made transcriptions always have an amount of uncertainty lying in the fact that this skill is hard to master [24]. While the percentage of uncertainty in phonetic transcriptions has not been measured to the best of our knowledge, we can assume that a deviation below 10% is enough to consider that a model-made transcription is similar to its human-made counterpart. We base this assumption on the statement that SLD detection tools are considered good when reaching an accuracy of over 80% [9].

Table 2. Metrics results for both Google-powered models compared with human-made results. Average deviations take human-made results as reference.

Transcription source	DAW	DRW	Average Accuracy	Average Deviation
Google API	86	20	86.05%	20.47%
Google-Phonemizer	95	18	75.16%	8.68%
Human-made	83	31	69.02%	N/A

We also noticed that the Google-Phonemizer model performs better at computing accuracies as close as possible to the human-made accuracies. This result was expected because this model produces phonetic transcriptions as opposed to the Google API model, which allows it to find more specific mistakes. Furthermore, it shows the reliability of Phonemizer for transcribing text to phonemes because the addition of this model to the accuracy pipeline resulted in a 11.79% decrease in deviation. It is worth to note that averages of accuracies are compared, which means that a same average does not imply identical accuracy values.

Indications on the number of DAW and DRW enables us to determine whether a model is more inclined to produce more or less words than the targeted amount, which was the human-made DAW and DRW values in this case. As shown in Table 2, all models tend to add more words instead of removing them, which is a similar tendency as in human-made transcriptions. This is mostly due to the tendency of readers to correct themselves when they struggle to pronounce a word. However, there are roughly 10 more transcriptions written by specialists that resulted in DRW. While analysing results row-by-row, we found that the phonetic precision in human-made transcriptions allowed them to highlight differences in similar phonemes (e.g. [ɔ] and [o]) and related to the child's accent (e.g. [ɛ] replaced by [e] in "j'ai") which is impossible to reproduce with Phonemizer since the French-Canadian accent is not supported yet. This leads to an increase of DRW since these kinds of differences generate an AW and a RW when detected.

Fig. 5 compares all three transcription sources by their number of AW and RW on average. Similar to the average deviation values, Google-Phonemizer is significantly closer to human-made transcriptions than Google-API on this regard. This difference is also related to the nature of the resulting transcription: while Google-API can only generate the closest word associated with a specific speech part, the other methods can represent more subtle changes in speech, like liaisons between words and accents.

Such information is useful for a SLP since each accent affects a specific range of phonemes that can be confused with pronunciation mistakes. However, while the accent is noticeable in human-made transcriptions and liaisons are part of phonetic models, differences caused by them have not been removed from the calculation of accuracy. Although the management of such exceptions was not in the scope of this work, it is worth to note that these problems will have to be addressed in order to make such systems more reliable in an end-to-end objective.

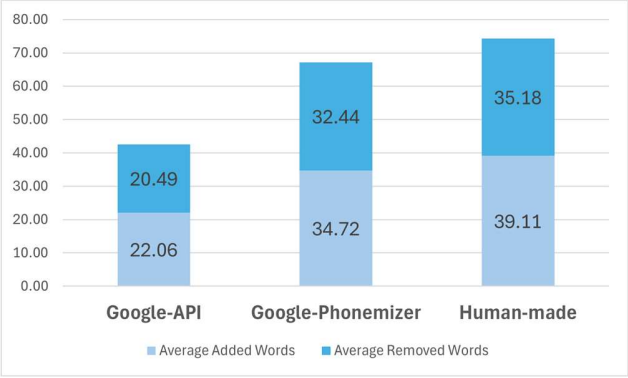

Fig. 5. Comparison of the average amount of added and removed words between Google API, Google-Phonemizer and human-made transcriptions.

VI. CONCLUSION AND FUTURE WORK

In conclusion, ASR models have shown a significant potential at helping SLPs to phonetically transcribe child speech for dyslexia diagnosis. We explored the usability of two model arrangements based on Google's speech-to-text API children's speech resembling evaluation scenarios and compared their outputs with transcriptions made by specialists. We concluded that even with off-the-shelf solutions, we can reach a deviation of less than 10% from human-made transcriptions, which is enough to consider a potential use in a clinical setting in the years to come.

We definitely want to further explore ASR tools that could improve the reported results, such as models trained on phonetically labelled speech. We also plan to expand our work on the dataset used in this study. While searching for data to carry out our experiments, we realized the lack of phonetically labelled speech in other languages than English for research in phonetics and ASR. The constitution of a French database of this kind is one of our future objectives to address this need.

ACKNOWLEDGMENT

We would like to thank FRQ-Inno for financially supporting this research. We would also like to express our gratitude to Cimon Chapdelaine for his significant contributions to the data collection and labeling process.

REFERENCES

[1] G. Schulte-Körne, "The Prevention, Diagnosis, and Treatment of Dyslexia," *Dtsch Arztebl Int*, vol. 107, no. 41, pp. 718–727, Oct. 2010, doi: 10.3238/arztebl.2010.0718.

[2] L. Rello and M. Ballesteros, "Detecting readers with dyslexia using machine learning with eye tracking measures," in *Proceedings of the 12th International Web for All Conference*, Florence Italy: ACM, May 2015, pp. 1–8. doi: 10.1145/2745555.2746644.

[3] T. Nevill and M. Forsey. (2023). "The social impact of schooling on students with dyslexia: A systematic review of the qualitative research on the primary and secondary education of dyslexic students". *Educational Research Review*, Volume 38, 100507.

[4] Y. Huang, M. He, A. Li, Y. Lin, X. Zhang and K. Wu. (2020). "Personality, Behavior Characteristics, and Life Quality Impact of Children with Dyslexia". *International Journal of Environmental Research and Public Health*, 17(4): 1415.

[5] B. Maughan, M. Rutter and W. Yule. (2020). "The Isle of Wight Studies: The Scope and Scale of Reading Difficulties". *Oxford Review of Education* 46, no. 4: 429-38.

[6] Z. Nurseitova and A. Shayakhmetova. (2023). Speech therapy to overcome dyslexia in primary schoolers. *Scientific Reports* 13, 4686.

[7] L. Marante, S. Hall-Mills and K. Farquharson. (2023). "School-Based Speech-Language Pathologists' Stress and Burnout: A Cross-Sectional Survey at the Height of the COVID-19 Pandemic". *Language, Speech, and Hearing Services in Schools*, Volume 54, 456-471.

[8] S. Bates, J. Watson, B. Heselwood and S. Howard. (2024). "Phonetic Transcription in Clinical Practice". *The Handbook of Clinical Linguistics*, Second Edition, Chapter 33.

[9] N. Albudoor and E. D. Peña. "Identifying Language Disorder in Bilingual Children Using Automatic Speech Recognition". *Journal of Speech, Language, and Hearing Research*, 1-14.

[10] PhonIA website. Avaliable at https://phonia.io

[11] A. Joshi, R. Bagate, Y. Hambir, A. Sapkal, N. P. Sable and M. Lonare. (2023). "System for Detection of Specific Learning Disabilities Based on Assessment". *International Journal of Intelligent Systems and Applications in Engineering*, 12(9s), 362–368.

[12] M. Alshehri, S. Sharma, P. Gupta and S. Ratan Shah. (2023). "Detection and Diagnosis of Learning Disabilities in Children of Saudi Arabia with Articial Intelligence". Preprint.

[13] L.M. Ward and Z. Kapoula. (2022). "Creativity, Eye-Movement Abnormalities, and Aesthetic Appreciation of Magritte's Paintings". *Brain Sci*, 12, 1028.

[14] L.M. Ward and Z. Kapoula. (2020). Differential diagnosis of vergence and saccade disorders in dyslexia. *Sci. Rep.*, 10, 22116.

[15] A.E. El Hmimdi, Z. Kapoula and V. Sainte Fare Garnot. (2024). "Deep Learning-Based Detection of Learning Disorders on a Large Scale Dataset of Eye Movement Records". *BioMedInformatics*, 4(1), 519-541.

[16] A.E. El Hmimdi, L.M. Ward, T. Palpanas, V. Sainte Fare Garnot and Z. Kapoula. (2022). "Predicting Dyslexia in Adolescents from Eye Movements during Free Painting Viewing". *Brain Sci.* 2022, 12, 1031.

[17] S. Parmar and C. Paunwala. (2023). "Early detection of dyslexia based on EEG with novel predictor extraction and selection". *Discov Artif Intell* 3, 33.

[18] E. Van Heuverswyn, C. Gosse and M. Van Reybroeck. (2024). "Handwriting difficulties in children with dyslexia: Poorer legibility in dictation and alphabet tasks, slowness in the alphabet task". *Dyslexia*, 0(2), e1767.

[19] G. Aldehim, M. Rashid, A. Saleh Alluhaidan, S. Sakri and S. Basheer. (2024). "Deep Learning for Dyslexia Detection: A Comprehensive CNN Approach with Handwriting Analysis and Benchmark Comparisons". *Journal of Disability Research*, Vol. 3(2).

[20] Quebec Order of Psychologists. (2014). "Lignes directrices pour l'évaluation de la dyslexie chez les enfants". Montreal, Canada.

[21] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang, Z. Meng, K. Hu, A. Rosenberg, R. Prabhavalkar, D. S. Park, P. Haghani, J. Riesa, G. Perng, H. Soltau, T. Strohman, B. Ramab-hadran, T. Sainath, P. Moreno, C.-C. Chiu, J. Schalkwyk,F. Beaufays, and Y. Wu. (2023). "Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages".

[22] M. Bernard and H. Titeux. (2021). "Phonemizer: Text to Phones Transcription for Multiple Languages in Python". *Journal of Open Source Software*, 6(68), 3958.

[23] R. H. Dunn and V. Vitolins. (2019). "eSpeak NG speech synthetizer". In *GitHub repository (Version 1.51), GitHub*.

[24] S. White, A. Hurren, S. James and R.-A. Knight. (2022). "'I think that's what I heard? I'm not sure': Speech and language therapists' views of, and practices in, phonetic transcription". *International Journal of Language and Communication Disorders*, 57(5): 1071-1084.