Applicability of Pretrained Language Models: Automatic Screening for Children's Language Development Level

Byoung-Doo Oh[1], **Yoon-Kyoung Lee**[2], **and Yu-Seop Kim**[1]
[1]Department of Convergence Software, Hallym University, Republic of Korea
[2]Division of Speech Pathology and Audiology, Hallym University, Republic of Korea
iambd822@gmail.com, {ylee, yskim01}@hallym.ac.kr

Abstract

The various potential of children can be limited by language delay or language impairments. However, there are many instances where parents are unaware of the child's condition and do not obtain appropriate treatment as a result. Additionally, experts collecting children's utterance to establish norms of language tests and evaluating children's language development level takes a significant amount of time and work. To address these issues, dependable automated screening tools are required. In this paper, we used pretrained LM to assist experts in quickly and objectively screening the language development level of children. Here, evaluating the language development level is to ensure that the child has the appropriate language abilities for his or her age, which is the same as the child's age. To do this, we analyzed the utterances of children according to age. Based on these findings, we use the standard deviations of the pretrained LM's probability as a score for children to screen their language development level. The experiment results showed very strong correlations between our proposed method and the Korean language test REVT (REVT-R, REVT-E), with Pearson correlation coefficient of 0.9888 and 0.9892, respectively.

1 Introduction

Language development is directly related to cognitive and intellectual development and is impacted by environmental factors including social interactions such as conversation with parents, etc (Sirbu, 2015). Language delay is the inability of a child to understand or use spoken language appropriately for their age, and it can result in language impairments. Language impairments are disorders of language that has a negative impact on all facets of life, including academic performance and social interaction, and restricts a child's wide range of potential (Bird et al., 1995; Conti-Ramsden and Botting, 2004; Hulme et al., 2020). In this situation, Tomblin et al. (1997) reported that many children with language impairment were not receiving appropriate treatment because their parents were unaware of the child's condition. In addition, many studies anticipate that following the COVID-19 pandemic, quarantine measures including social distancing and mask wearing will include a negative impact on children's language development (Charney et al., 2021; Deoni et al., 2021; Viola and Nunes, 2022).

To address this issue, experts have developed language tests that may be used prior to make diagnosing language impairments. Standardized formal test analyzes linguistic abilities to screening a child's language development level. For example, PPVT-IV (Peaboby Picture Vocabulary Test-IV) (Dunn and Dunn, 2006) and EVT-2 (Expressive Vocabulary Test-II) (Kathleen T. Williams, 2008) evaluate receptive vocabulary and expressive vocabulary, respectively. Language sample analysis (LSA) analyzes linguistic abilities like grammar, pragmatics, and semantics as a measure (Schober-Peterson and Johnson, 1993; Robert E. Owen Jr, 2013). These methods evaluate a child's language development level compared with standardized norms from the same age group's children who have normally developed. In other words, it evaluates if a child has linguistic abilities that are age-appropriate. If a child's scores on these methods are lower than the norm for the same age group, tests to diagnose language impairment are performed. However, moving forward with the standardized formal test and LSA process requires

Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI), pages 149 - 156
December 7, 2022

Topic	Family			
Turn	**Number**	**Person**		**Utterances**
		Interviewer	KR	어제 형이랑 뭐하고 놀았어?
			EN	What did you play with brother yesterday?
1	1	Child	KR	(장난감) 장난감 가지고 놀고 청소도 했어요.
			EN	We played with (toys) toys and cleaned.
	2	Child	KR	그리고 (음) 형이 자꾸 나만 시켜요.
			EN	And (um) my brother keeps making me do it.
		Interviewer	KR	아 그랬구나.
			EN	Oh, I see.
		…		…

Table 1: Example of the data collected by the Hallym Conversation & Pragmatic Assessment Protocol.

a lot of time and work, and the same is true for establishing reliable standardized norm.

Consequently, recent studies tried to an automated screening test that used the acoustic features of children's speech (Maier et al., 2009; Gong et al., 2016). They classified children with speech and language impairments from those with typical development using machine learning which is support vector machine and linear regression. They made it easier to collect data and made it possible to develop a system that could automatically screen for children's speech and language impairments. However, it still has to depend on data to train machine learning models, and cannot be used in another languages. At the same time, they only classified normal and impaired, and it is difficult to distinct the language development level like the existing language tests. In particular, although acoustic features are suitable for discriminating speech impairments due to problems such as speech organs, it is not suitable for discriminating language impairments because it has no linguistic characteristics.

The pretrained language model (pretrained LM), such as GPT2 (Radford et al., 2019) and GPT3 (Brwon et al., 2020), is being developed for a variety of languages and has achieved good performance in a variety of downstream tasks of natural language processing. In the grammatical error correction (GEC) task, studies using only pretrained LM have been performed (Bryant and Briscoe, 2018; Yasunaga et al., 2021). To identify grammatical errors in sentences, Bryant and Briscoe (2018) and Yansunaga et al. (2021) used normalized log probability and probability score, respectively, based on the pretrained LM. The basis for these studies was the observation that grammatical sentences (s_{good}) had a higher probability score of the pretrained LM than non-grammatical sentences (s_{bad}).

$$p(s_{bad}) < p(s_{good}) \quad (1)$$

Based on these characteristics, we focused on a pretrained LM's applicability like unsupervised learning that do not depending on training data for a specific task. In this paper, we used pretrained LM to assist experts in quickly and objectively screening the language development level of children. First, the pretrained LM calculates the probability of a word sequence for each utterance (i.e. sentence) of the child. Following that, a screening score for children's language development level is calculated using the standard deviation of these scores. The advantages of this method are as follows:

- Since it doesn't need procedures like fine-tuning carried out in supervised learning, it doesn't depend on data. As a result, it is relatively free of the cost and time required for data collection.
- It can screen not only children whose language development is slow, but also children whose language development is fast.
- It can be applied in various languages differently from another automated screening methods because pretrained LMs are being developed for various languages.

The format of this paper is as follows. The data we used are described in Section 2. Section 3 describes how to screen children's language development level using pretrained LMs. The

Age	Children	No. of Sentences (Avg)	No. of Tokens
2-year-old	16	69.13	3K
3-year-old	17	104.44	6K
4-year-old	43	89.34	21K
5-year-old	40	83.93	21K
6-year-old	27	102.85	20K
Total	143	89.94	71K

Table 2: Details on our age-specific data.

Type	Details
Maze words	Repetitions Revisions
Silence pauses	More than 3 seconds
Inaccurate pronunciation	-

Table 3: Removed word types.

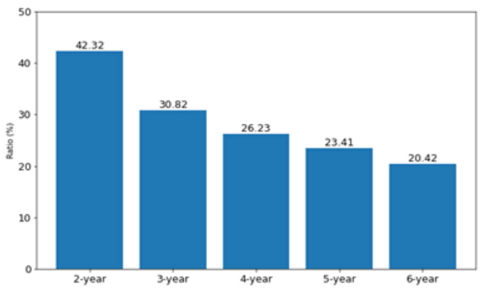

Figure 1: Ratio of age-specific single-word utterances.

experimental settings and results are discussed in Section 4, and our findings and conclusions are compiled in Section 5. Finally, we discuss the ethical considerations and limitations of our proposed method in Section 6.

2 Transcription Data

In the field of speech therapy, a rule called conversation protocol is used to ensure reliability of norms and analysis when collecting children's data (i.e. utterances). Conversation protocols allow only specific topics in the interview, and experts encourage children to speak on their own. As a result, we used the Hallym Conversation & Pragmatic Assessment Protocol (Lee and Choi, 2017) that the Division of Speech Pathology and Audiology at Hallym university created to collect data. Standardized formal test and LSA examine the age-related differences in scores in children who have developed normally in order to verify their norms. So, children between the ages of 2 and 6 who the experts assessed to have developed normally were the subjects of data. The data we used were collected by experts after approved the Institutional Review Board of Hallym University[1]. It is a total of 143 children, and each child includes an average of 89 utterances. Table 1 shows the collected data, while Table 2 shows age-specific details of the data.

The words that are used by people of all ages and make it interrupt to analyze a language development level are indicated with special characters (i.e. symbols) in LSA and then excluded from the analysis. As a result, after analyzing them, we removed these words. These word types are shown in the following Table 3.

Single-word utterances, such as “yes” or “no” and utterances using only proper nouns for people or things, are frequently appeared in children. Because these utterances frequently appear in children of all ages, they interfered with the classification of children's age for language development levels in previous studies based on supervised learning (Oh et al., 2021; Oh et al., 2022). As children grow older, these utterances tend to become less frequent and utterances with complete sentence become more. We believe that this tendency is a useful linguistic characteristic for screening to children's language development level based on pretrained LM. The ratio of these utterances in age-specific children's overall utterances is shown in Figure 1.

[1] The study was conducted according to the guidelines of the Declaration of Helsinki and is approved by the Institutional Review Board of Hallym University (HIRB-2019-036, HIRB-2021-093).

LM	Params	Ratio of $p(s_{bad}) < p(s_{good})$
KoGPT2-SKT	125M	72.7 %
KoGPT3-Kakao	6B	83.2 %

Table 4: Correlation with grammar assessment for Korean pretrained LM's probability.

3 Automatic Screening based on Pretrained LMs

Children's language systems, including their grasp of grammar, steadily improve as they grow older. In this situation, the basis of our proposed method can evaluate the sentence's grammaticality using several values (e.g., probability, normalized log probability, etc) that can be calculated from a pretrained LM (Bryant and Briscoe, 2018; Yasunaga et al., 2021). These characteristics demonstrate the feasibility of using a pretrained LM to screen children's language development level. To screen the child's language development level, we only use the pretrained LM's probability as a score for the child's utterance and calculate the standard deviation of these scores. The rest of this section details more into pretrained LMs which is used in this paper and the scoring method we used to screen for children's language development level.

3.1 Pretrained LMs for Korean

Yasunaga et al. (2021) verified that the pretrained LM's probability may be used to assess the grammaticality of English sentences. Based on GPT2, grammatical sentences (s_{good}) were evaluated highly scores in around 94% of all the data which is consist of (s_{good}, s_{bad}) pairs. So, we verified if the Korean pretrained LM provided the same observations as these experiments.

In this paper, we used KoGPT2[2] (KoGPT2-SKT) released by SK Telecom Co., Ltd and KoGPT3[3] (KoGPT3-Kakao) released by Kakao Corp. as Korean pretrained LMs. Additionally, we used the Korean grammaticality assessment corpus (National Institute of Korean Language) to validate the Korean pretrained LMs. The Korean grammaticality assessment corpus consists of (s_{good} , s_{bad}) pairs. Korean pretrained LMs likewise had the same tendency as the observations of Yasunaga et al. (2021), as shown in Table 4.

3.2 Scoring for language development level

We evaluate the language development level with all utterances the child makes in conversations with experts. Consequently, the score for the utterance was calculated by the probability of a word sequence in the pretrained LM.

$$p(s_i) = P(w_1, w_2, w_3, \dots, w_n) \tag{2}$$

$$p(child) = [p(s_1), p(s_2), \dots, p(s_i)] \tag{3}$$

, where s_i is the i-th utterance and w_n is the n-th word that makes up s_i, $p(s)$ is a score for one utterance, and $p(child)$ is score set calculated for child's all utterances.

However, these scores can be verified as in Equation (1) only by comparing s_{good} and s_{bad} having the same meaning. And the data we used was collected by having a conversation about a specified topic, however these topics have a wide meaning such as family and friend. We may organize these issues into the following three intuitions:

Intuition (1). Relativity of probability distributions for sentences. A grammatical sentence gets a high score based on the pretrained LM's probability. It can evaluate grammaticality in sentences that have the same meaning. As a result, even though they are grammatical sentences, sentences with different meanings have different probability distributions.

Intuition (2). A conversational topic having a wide meaning. Each child might have a different story to tell even about the same topic because the topic is so broad. For example, while talking friends, child-A can talk a story he played with friend, hereas child-B can talk a story about a conflict with friend. In other words, the utterances' contents differ from one another.

Intuition (3). Age-related variations in the frequency of single-word utterances. As shown in Section 2, children use basic positive and negative words like "yes" and "no" less frequently as they grow older. That is, people of all ages use these words.

These intuitions can be summed up as follows: children's utterances have different probability

[2] https://github.com/SKT-AI/KoGPT2

[3] https://github.com/kakaobrain/kogpt

Methods	Age group				
	2-year-old	**3-year-old**	**4-year-old**	**5-year-old**	**6-year-old**
REVT-R	18.04	30.35	44.39	58.18	70.92
REVT-E	20.16	37.06	52.38	64.81	75.06
Ratio of single-word utterances	29.25	19.65	23.43	18.73	21.0
KoGPT2-SKT	12.97	18.43	30.50	34.11	41.55
KoGPT3-Kakao	12.02	16.06	26.84	30.09	36.26
Correlation Coefficient (r)	Ratio of single-word utterances	REVT-R		-0.6451	
		REVT-E		-0.6946	
	KoGPT2-SKT	**REVT-R**		**0.9888**	
		REVT-E		**0.9892**	
	KoGPT3-Kakao	REVT-R		0.9876	
		REVT-E		0.9868	

Table 6: Experiment results of the correlation analysis for our proposed method and REVT.

Age	Average	Max	Min
2-year-old	3.14	18	1
3-year-old	3.88	38	1
4-year-old	5.49	108	1
5-year-old	6.43	72	1
6-year-old	7.27	85	1

Table 5: Details on token length in age-specific sentence.

distributions. For instance, single-word utterances will get a lower score. Additionally, even when speaking on the same topic in complete sentences, the distribution of scores may differ. To utilize pretrained LM's probability correctly, we must get around these limits. We believed that characteristics of linguistic which is universal and changes with age-specific, it may be a key in overcoming these limits. We concluded that the solution is a departure from the single-word utterances that always emerges inside different probability distributions, which can be summarized as follows: (1) The deviation of probability is little since single-word utterance occurs more frequently as the child becomes younger. (2) As children grow older, the deviation of probability is bigger since single-word utterances and utterances with complete sentence appearing appropriately. Consequently, to screen the children's language development levels, we calculated the standard deviation of the $p(child)$.

$$score(child_N) = \sqrt{\frac{(p(s_1)-\mu)^2 + \ldots + (p(s_i)-\mu)^2}{i}} \quad (2)$$

, where μ is the average of the pretrained LM's probability for the child's utterances and i is the number of utterances.

4 Results and Discussion

To ensure consistency and reliability of the analysis, LSA chooses 30 to 50 of the utterances made by children and analyzes them as a certain number of utterances (Harris et al., 1986; Ingram, 2002; Trudeau and Sutton, 2011; Andonova, 2015). By omitting this procedure, we aim to provide an automated screening that experts can use easily and quickly. As a result, we evaluated by the child's all utterances. This data, which is detailed in Table 5, includes utterances of various lengths.

REVT (Hong et al., 2009) is a standardized formal test in the Korean language that measures both receptive (REVT-R) and expressive (REVT-E) vocabulary in individuals between the ages of 2 and 16. The norms of REVT-R and REVT-E were constructed by the Seoul Community Rehabilitation Center for the disabled to children who have normally developed of 5,119 and 5,145 individuals, respectively, and provided for use. Consequently, to evaluate the reliability of our proposed method, we evaluated the correlation with the norms of the REVT. The results as shown in Table 6.

Table 6 shows the standardized norms or calculated scores for which each method by age. First, we confirmed whether a simple method, ratio of single-word utterances, could be used as the age-specific score for children's language development level. This is because we confirmed that there was a significant difference by age in Figure 1. But it showed a very low correlation with REVT. Our investigation revealed that the reason was that some children their age used single-word utterance more frequently. Next, we confirmed the possibility of our proposed method. It was able to confirm a strong correlation with REVT. KoGPT2-SKT in particular shown extremely strong correlation with REVT-R and REVT-E, with Pearson correlation coefficients of 0.9888 and 0.9892, respectively. Despite being little less than this, KoGPT3-Kakao also showed a respectable correlation. In actuality, KoGPT3-Kakao is a latest model, and as shown by Table 4, it performs better in grammar assessment. We believe that the somewhat different model structures in the two pretrained LM—as well as the different training dataset—are what caused the difference in the correlation coefficients. These findings demonstrated the potential for using a pretrained LM to address the limitations of language tests, which are expensive, time-consuming, and difficult to utilize across a variety of languages.

5 Conclusion

In this paper, we used pretrained LMs for automated screening and tried to address limitations in the existing language tests, such as the number of data and the diversity of languages. At this time, we preprocessed the utterance by analyzing age-specific linguistic patterns of children to use the pretrained LM efficiently. Additionally, the correlation with REVT, a standardized formal test for Korean language, was evaluated to demonstrate the reliability of our proposed method. The experimental results revealed a strong correlation between our proposed method, which is based on KoGPT2-SKT, and the norms for REVT-R and REVT-E, with Pearson correlation coefficients of 0.9888 and 0.9892, respectively. These observations demonstrate the potential for the pretrained LM to automatically screen children's language development levels and are expected to address several issues with the limitations of language tests such as standardized formal tests and LSA.

Furthermore, we believe that the pretrained LM demonstrated the potential for applicability in various issues needing skills in natural language processing. Future work will focus on make up for automatic screening based on pretrained LMs and investigating automatic transcription methods for collecting children’s utterance data using automatic speech recognition.

6 Ethical Considerations and Limitations

If our proposed method is successful, it is possible to screen a child’s language development level quickly and objectively prior to having an expert perform a language test. And if a problem is identified at this time, the child can get early diagnostic tests and treatment. Additionally, because expert direct analysis is not included, the language test's cost may be reduced, increasing its accessibility to parents. As the language test gets easier, though, it's possible that unneeded diagnoses and treatments may be provided.

Next, the issues that could occur if our proposed method operates improperly were then taken into consideration as follows: The first is the failure to screening for children who has abnormally developed (recall failure). Recall failure has a problem of missing the treatment time because it cannot properly diagnose and treat a child who has abnormally developed. Second, it involves screening children who has normally developed (precision failure). To children who has normally developed, precision failure can lead in unneeded diagnosis and treatment. We also take into consideration the following potential misuses of this method: Future issues with discrimination might arise if this method is expanded to evaluate children's intellectual development level. In other words, it is possible to discriminate and educate children with high and low developmental levels, which undermining the fundamental purpose of education. Consequently, this method should be performed under strictly managed by a group of experts in relevant fields, such as language pathology or speech therapists.

Technically, the method we propose relies solely on a pretrained LMs; no extra learning, such as fine-tuning, is involved. Consequently, this method is relatively free to the bias issue that training data in supervised learning might bring. The bias of the corpus that was used to develop the pretrained LM at this time may cause some concern. However, the appropriacy and factuality of a sentence's content are not factors we believe should be taken into consideration when evaluating a child's language development level. And, since this technique does not need for extra training, it does not consider the data collection from users. Although we cannot collect it directly right now since speech recognition technique is not being employed, but this will change as technique advances. Consequently, these applications must adhere to research ethics regulations such as the IRB for data collection.

Finally, the test results of our proposed method, including the language test, may vary depending on the level of participation like the child's sociable or active nature. Consequently, we have to take these into consideration as well.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A5A8019303), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MIST) [No.2021-0-02068, Artificial Intelligence Innovation Hub (Artificial Intelligence Institute, Seoul National University)], and ´R&D Program for Forest Science Technology (Project No. 2021390A00-2123-0105)´ provided by Korea Forest Service (Korea Forestry Promotion Institute).

References

Elena Andonova. 2015. Parental report evidence for toddlers' grammar and vocabulary in Bulgarian. *First Language*, 35(2):126-136.

Judith Bird, Dorothy VM. Bishop, Norman H. Freeman. 1995. Phonological awareness and literacy development in children with expressive phonological impairments. *Journal of Speech, Language, and Hearing Research*, 38(2):446-462.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *Advances in neural information processing systems*, 33, pages 1877-1901.

Christopher Bryant and Ted Briscoe. 2018. Language model based grammatical error correction without annotated training data. In *Proceedings of 13th workshop on innovative use of NLP for building educational applications*, pages 247-253, New Orleans, Louisiana. Association for Computational Linguistics.

Sara A. Charney, Stephen M. Camarata, Alexander Chem. 2021. Potential impact of the COVID-19 pandemic on communication and language skills in children. *Otolaryngology–Head and Neck Surgery*, 165(1):1-2.

Gina Conti-Ramsden and Nicola Botting. 2004. Social difficulties and victimization in children with SLI at 11 years of age. *Journal of Speech, Language, and Hearing Research*, 47(1):145-161.

Sean CL. Deoni, Jennifer Beauchemin, Alexandra Volpe, Viren D'Sa, Resonance Consortium. 2021. Impact of the COVID-19 pandemic on early child cognitive development: initial findings in a longitudinal observational study of child health. MedRxiv:2021.08.10.21261846. Version 2.

Lloyd M. Dunn and Duglas M. Dunn. 2007. *Peabody Picture Vocabulary Test Fourth Edition*. Bloomington, MN: NCS Pearson Inc.

Jen J. Gong, Maryann Gong, Dina Levy-Lambert, Jordan R. Green, Tiffany P. Hogan, John V. Guttag. 2016. Towards an Automated Screening Tool for Developmental Speech and Language Impairments. In *Proceedings of 17th Annual Conference of the International Speech Communication Association*, pages 112-116.

Margaret Harris, David Jones, Susan Brookes, Julia Grant. 1986. Relations between the non-verbal context of maternal speech and rate of language development. *British journal of developmental psychology*, 4(3):261-268.

Gyung-Hun Hong, Young-Tae Kim, Kyung-Hee Kim. 2009. Content and Reliability Analyses of the Receptive and Expressive Vocabulary Test (REVT). *Korean Journal of Communication Disorders*, 14(1):34-45. [in Korean].

Charles Hulme, Margaret J. Snowling, Gillian West, Arne Lervåg, Monica Melby-Lervåg. 2020. Children's language skills can be improved: Lessons from psychological science for educational policy. *Current Directions in Psychological Science*, 29(4):372-377.

David Ingram. 2002. The measurement of whole-word productions. *Journal of Child Language*, 29(4):713-733.

Yoon-Kyoung Lee and Ji-Eun Choi. 2017. *Hallym conversation and pragmatic assessment protocol*. Manuscript in preparation.

Andreas M. Maier, Tino Haderlein, U Eysholdt, Frank Rosanowski, Anton Batliner, Maria E. Schuster, Elmar Nöth. 2009. PEAKS–A system for the automatic evaluation of voice and speech disorders. *Speech Communication*, 51(5):425-437.

National Institute of Korean Language. 2021. *National Institute of the Korean Language Grammaticality Assessment Corpus*, Version 1.1. [in Korean].

Byoung-Doo Oh, Yoon-Kyoung Lee, Hye-Jeong Song, Jong-Dae Kim, Chan-Young Park, Yu-Seop Kim. 2021. Age group classification to identify the progress of language development based on convolutional neural networks. *Journal of Intelligent & Fuzzy Systems*, 40(4):7745-7754.

Byoung-Doo Oh, Yoon-Kyoung Lee, Jong-Dae Kim, Chan-Young Park, Yu-Seop Kim. 2022. Deep Learning-Based End-to-End Language Development Screening for Children using Linguistic Knowledge. *Applied Sciences*, 12(9):4651-4664.

Robert E. Owen Jr. 2013. *Language disorders: A functional approach to assessment and intervention 6th Ed*. Allyn and Bacon, Boston, MA.

Debra Schober-Peterson and Cynthia J. Johnson. 1993. The performance of eight-to ten-year-olds on measures of conversational skilfulness. *First Language*, 13(38):249-269.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9.

Anca Sirbu. 2015. The significance of language as a tool of communication. *Scientific Bulletin" Mircea cel Batran" Naval Academy*, 18(2):405.

Bruce J. Tomblin, Nancy L. Records, Paula Buckwalter, Xuyang Zhang, Elaine Smith, Marlea O'Brien. 1997. Prevalence of specific language impairment in kindergarten children. *Journal of speech, language, and hearing*, 40(6):1245-1260.

Natacha Trudeau and Ann Sutton. 2011. Expressive vocabulary and early grammar of 16- to 30-month-old children acquiring Quebec French. First Language, 31:480-507.

Thiago Wendt Viola and Magda Lahorgue Nunes. 2022. Social and environmental effects of the COVID-19 pandemic on children. *Jornal de pediatria*, 98:4-12.

Kathleen T. Williams. 2007. *Expressive Vocabulary Test Second Edition*. Circle Pines, MN: AGS Publishing.

Michihiro Yasunaga, Jure Leskovec, Percy Liang. 2021. LM-Critic: Language Models for Unsupervised Grammatical Error Correction. In *Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7752-7763, Online and Punta, Dominican Republic. Association for Computational Linguistics.