IberSPEECH 2024
11-13 November 2024, Aveiro, Portugal

iRead4Skills @ IberSPEECH 2024: Project presentation and developments for the Portuguese language

Jorge Baptista[1,2], *Eugénio Ribeiro*[1,3], *Nuno Mamede*[1,4]

[1]INESC-ID Lisboa, Portugal [2]Universidade do Algarve, Portugal [3]Instituto Universitário de Lisboa (ISCTE-IUL), Portugal [4]Instituto Superior Técnico, Universidade de Lisboa, Portugal

jorge.baptista@inesc-id.pt, eugenio.ribeiro@inesc-id.pt, nuno.mamede@inesc-id.pt

Abstract

The **iRead4Skills** project aims to enhance adult literacy and essential skills by merging technology and culture. It addresses the needs of adult learners, particularly those with low literacy skills, by providing an intelligent system that evaluates text complexity and suggests readings suited to individual levels. This open-access system supports multilingual environments, offering resources in languages like Portuguese, Spanish, and French. The project innovates by using end-user input to develop new text complexity measures, thus aligning learning tools more closely with real-world literacy needs. It also emphasizes the role of technology, especially in the area of NLP, in tailoring educational materials for trainers and learners. Through collaboration with various stakeholders — including universities, government bodies, and research institutions — the project aims to inform policymakers and educators about ways to improve workforce skills and foster lifelong learning across Europe.

Index Terms: adult literacy, text complexity evaluation, multilingual support, digital literacy tools, workforce skills development

1. Introduction

The **iRead4Skills**[1] project is an on-going innovative initiative aimed at enhancing adult literacy [1, 2, 3] by leveraging advanced technologies, including Human Language Technologies (HLT) [4, 5, 6, 7]. Recent research in this area focuses on the application of Artificial Intelligence (AI) approaches, resourcing not only to traditional Machine Learning (ML) techniques [8, 9] but also to the Deep Learning (DL) methods that are currently prominent in Natural Language Processing (NLP) [10, 11]. The project's core goal is to develop an intelligent system that evaluates the complexity of texts and suggests appropriate readings based on the literacy level of adult learners [12, 13, 14]. Complexity levels for low-literacy adult native speakers can be defined [15] based three text complexity levels, aligned with the Common European Framework of Reference for Languages (CEFR) [16] — *Very Easy* (corresponding to A1), *Easy* (corresponding to A2), and *Plain/Clear* (corresponding to B1). The system is also designed to identify complex structures within texts and offer recommendations for enhancing readability, ensuring that the content aligns with the literacy skills of the intended audience. By focusing on improving both individual literacy and workforce skills, the project contributes to societal well-being and supports learners in adapting to the rapidly changing demands of the labor market [17].

[1]https://iRead4Skills.com/ Grant: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837.

The **iRead4Skills** project, led by Universidade Nova de Lisboa (UNL, Portugal)[2], is a collaboration among multiple institutions across Europe. These include the Ministry of Education (Portugal)[3], the Autònoma University of Barcelona (UAB, Spain)[4], the Catholic University of Louvain (UCL, Belgium)[5], the University of Santiago de Compostela (USC, Spain)[6], the Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento em Lisboa (INESC-ID Lisboa)[7], the Luxembourg Institute of Socio-Economic Research (LISER, Luxembourg) [8], and Mindshaker (Portugal)[9]. Several stakeholders are involved in the project development. These include the Agência Nacional para a Qualificação e o Ensino Profissional (ANQEP, Portugal)[10], the European Basic Skills Network (EBSN),[11] the Swiss Federation for Adult Learning (SVEB/FSEA, Switzerland)[12] the Agence National de Lutte contre l'Illetrisme (France)[13] the Association Savoirs pour Réussir Paris [14], and the Département d'Ille-et-Vilaine [15].

INESC-ID Lisboa and, in particular, its HLT lab[16], plays a key role for the project's scientific and technological goals, with its expertise on the processing of the Portuguese language [18, 19, 20, 21] paired with recent developments in NLP and AI in general [11, 22]. This paper provides an overview of the project and highlights the role of the INESC-ID team and its focus on developments for the Portuguese language.

2. Project Overview

The **iRead4Skills** project aims to provide a platform that supports adult learners and educators by developing tools for text evaluation and adaptation. The system is multilingual, with support for languages such as Portuguese [11], Spanish, and French [23], making it accessible to a broad audience. One of the key features is its intelligent text evaluation capability, which analyzes the complexity of a given text [15] and recommends improvements that can match the user's proficiency level.

The project focuses on aligning literacy programs with the

[2]https://www.unl.pt
[3]https://www.sec-geral.mec.pt/
[4]https://www.uab.cat/
[5]https://uclouvain.be/
[6]https://www.usc.gal
[7]https://www.inesc-id.pt/
[8]https://www.liser.lu/
[9]https://mindshaker.com/
[10]https://www.anqep.gov.pt/
[11]https://basicskills.eu/
[12]https://alice.ch/
[13]https://www.anlci.gouv.fr/
[14]https://sprparis.wordpress.com/
[15]https://www.ille-et-vilaine.fr/
[16]https://www.hlt.inesc-id.pt/

10.21437/IberSPEECH.2024-62

needs of the labor market [14]. By fostering digital literacy and improving reading skills, the system helps workers adapt to new technologies and job market demands. It also supports lifelong learning, emphasizing the role of cognitive and socioemotional skills in workforce success.

The project provides access to a variety of project outcomes, including reports, datasets, and tools developed throughout the initiative. Key resources include surveys on reading skills and workforce gaps, literature reviews on readability and complexity features [24], basic lexicons per complexity level [25], and annotated corpora categorized by text complexity in multiple languages (Portuguese, Spanish, and French) [26]. These documents aim to support educators, policymakers, and researchers in enhancing literacy and reading comprehension across different domains. Additionally, tools like annotation manuals and lexicons for complexity levels [27] are available to facilitate text evaluation and adaptation for diverse proficiency levels. This section reflects the project's commitment to improving adult reading skills and bridging literacy gaps.

3. The Role of INESC-ID Lisboa

The INESC-ID team focuses on the computational processing of the Portuguese language and contributes to two major scientific aspects of the project. First, from a linguistic perspective, the team focuses on determining features of textual complexity grounded in textual evidence [28, 29, 30, 31]. This work aligns with the CEFR [16, 32, 33] and supports language learning, text evaluation, and related domains [34]. Specifically, the team investigates how various linguistic features, such as syntax [35], vocabulary richness [36, 37], and sentence structure [35], correlate with different CEFR proficiency levels, thereby enabling more precise text evaluation [38].

From a computational perspective, the INESC-ID team explores the use of several NLP techniques to analyze text complexity. The team addresses several key problems in this area by combining modern approaches based on DL, foundation models, and Large Language Models (LLMs) with traditional ML models, to promote explainability in the results. By integrating these techniques, the team aims to provide insights that can be easily interpreted by educators and learners, facilitating more effective text adaptation and language instruction.

The combination of linguistic expertise and computational advancements makes INESC-ID Lisboa a key contributor to the success of the **iRead4Skills** project. The team's work on text complexity assessment not only improves the accuracy of the system's recommendations but also enhances its adaptability across different educational contexts and languages.

4. Scientific Goals

The specific scientific goals of the INESC-ID team are twofold: **Linguistic Perspective:** The team aims to determine the features of textual complexity based on linguistic evidence. This involves studying the relationship between linguistic characteristics of texts and language proficiency, particularly within the low-literacy range, and relating it to the CEFR [16]. By identifying these features, the team supports the development of tools for evaluating and adapting texts for different proficiency levels. **Computational Perspective:** The team focuses on integrating DL models [39, 40, 10, 41] with traditional ML techniques to analyze text complexity. By combining these approaches, they aim to achieve some level of explainability in the results, making the system's recommendations more transparent and useful for educators and learners.

5. Impact and Future Directions

The **iRead4Skills** project has the potential to significantly impact adult literacy education in Europe. By providing tools for intelligent text evaluation and adaptation, the project enables educators to better tailor their materials to the needs of learners, particularly those with lower literacy levels. The project's emphasis on using technology to support language learning aligns with broader European efforts to improve digital literacy and workforce skills.

Looking ahead, the INESC-ID team plans to expand its research on text complexity and explore further applications of foundation models and LLMs in the context of language education. Additionally, the team aims to refine the explainability of DL techniques to ensure that the system's recommendations are interpretable and avoid the black-box methodologies currently governing a large part of NLP research. These efforts will contribute to the project's overarching goal of fostering more inclusive access to literacy and education across Europe.

6. Conclusion

The **iRead4Skills** project represents a significant step forward in addressing the literacy challenges faced by adults in Europe. Through its intelligent text evaluation system and multilingual support, the project enables learners to improve their reading skills and adapt to the changing demands of the workforce. INESC-ID Lisboa and, particularly, its HLT lab plays a key role in this effort, contributing both linguistic and computational expertise to enhance the project's impact.

7. Acknowledgements

This work was partially supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (Reference: UIDB/50021/2020, DOI: 10.54499/UIDB/50021/2020) and by the European Commission (Project: **iRead4Skills**, Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837).

8. References

[1] A. Steeds, *Adult Literacy Core Curriculum Including Spoken Communication*, 2001.

[2] M. J. Alves and S. Lameira, *Referencial de Competências-chave de Educação e Formação de Adultos – Nível Básico*. ANQEP, I.P., 2021.

[3] F. Clément, L. Hauret, R. Amaro, and R. Duarte, "iRead4Skills - Literature Review & Report on Literacy and other Skills," Jul. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.13127361

[4] K. Collins-Thompson, "Computational assessment of text readability: A survey of current and future research," *International Journal of Applied Linguistics*, vol. 165, no. 2, pp. 97–135, 2014.

[5] S. A. Crossley, S. Skalicky, M. Dascalu, D. S. McNamara, and K. Kyle, "Predicting text comprehension processing and familiarity in adult readers: New approaches to readability formulas," *Discourse Processes*, vol. 54, no. 5-6, pp. 340–359, 2017.

[6] S. A. Crossley, K. Kyle, and M. Dascalu, "The Tool for the Automatic Analysis of Cohesion 2.0," *Behavior Research Methods*, vol. 51, no. 1, pp. 14–27, 2019.

[7] S. A. Crossley, S. Skalicky, and M. Dascalu, "Moving beyond classic readability formulas: New methods and new models,"

Journal of Research in Reading, vol. 42, no. 3-4, pp. 541–561, 2019.

[8] S. Akef, A. Mendes, D. Meurers, and P. Rebuschat, "Investigating the Generalizability of Portuguese Readability Assessment Models Trained Using Linguistic Complexity Features," in *PROPOR 2024*, 2024, pp. 332–341.

[9] X. Chen and D. Meurers, "CTAP: A Web-based Tool Supporting Automatic Complexity Analysis," in *CL4LC*, 2016, pp. 113–119.

[10] B. W. Lee, Y. S. Jang, and J. Lee, "Pushing on text readability assessment: A transformer meets handcrafted linguistic features," in *EMNLP 2021*, 2021, pp. 10 669–10 686.

[11] E. Ribeiro, N. Mamede, and J. Baptista, "Automatic Text Readability Assessment in European Portuguese," in *PROPOR 2024*, 2024, pp. 97–107.

[12] OECD, "Technical Report of the Survey of Adult Skills (PIAAC)," OECD Publishing, Paris, Tech. Rep., 2013.

[13] ——, *The Survey of Adult Skills Reader's Companion*, Paris, 2013.

[14] UNESCO, "Literacy," 2020, available at: https://uis.unesco.org/node/3079547.

[15] R. Monteiro, R. Amaro, S. Correia, A. Pintard, R. Gauchola, M. Moutinho, and X. Blanco Escoda, "iread4skills - complexity levels," Jan. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.10459090

[16] Council of Europe, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Strasbourg: Council of Europe Publishing, 2020. [Online]. Available: https://www.coe.int/lang-cefr

[17] R. Amaro, S. Correia, C. Gramacho, and A. Mendes, "Automatização no diagnóstico de nível de língua: anotação e versatilidade dos recursos para PLE," *Revista da Associação Portuguesa de Linguística*, no. 7, p. 1–20, 2020. [Online]. Available: https://ojs.apl.pt/index.php/rapl/article/view/86

[18] N. Mamede, J. Baptista, V. Cabarrão, and C. Diniz, "STRING: A hybrid statistical and rule-based Natural Language Processing chain for Portuguese," in *International Conference on Computational Processing of Portuguese (PROPOR 2012)*, vol. Demo Session, Coimbra, Portugal, April 2012.

[19] R. Correia, J. Baptista, N. Mamede, I. Trancoso, and M. Eskenazi, "Automatic Generation of Cloze Question Distractors," in *SLaTE 2010, Second Language Studies: Acquisition, Learning, Education and Technology*. Waseda University, Tokyo, Japan (September 22, 2010): Interspeech 2010/ISCA SIG on Speech and Language Technology in Education, 2010.

[20] P. Curto, N. Mamede, and J. Baptista, "Automatic Text Difficulty Classifier," in *CSEDU 2015*, vol. 1, 2015, pp. 36–44.

[21] N. Mamede, J. Baptista, I. Trancoso, A. Ferreira, R. Correia, A. Silva, and C. Marques, "Reap.pt - an online portuguese tutor," http://www.propor2012.org/demos/DemoREAP_PT.pdf, 2012.

[22] E. Ribeiro, N. Mamede, and J. Baptista, "Text Readability Assessment in European Portuguese: A Comparison of Classification and Regression Approaches," in *PROPOR 2024*, 2024, pp. 551–557.

[23] R. Wilkens, D. Alfter, X. Wang, A. Pintard, A. Tack, K. P. Yancey, and T. François, "FABRA: French aggregator-based readability assessment toolkit," in *ELRA*, 2022, pp. 1217–1233.

[24] A. Pintard, T. François, J. Baptista, M. G. González, R. Wilkens, E. Ribeiro, and R. Amaro, "Work package 4: Report on readability features," iRead4Skills Project, Tech. Rep., October 2023, internal technical report.

[25] R. Wilkens, A. Pintard, T. François, S. Barbosa, M. L. Reis, R. Amaro, E. Ribeiro, N. Mamede, J. Baptista, X. Blanco, A. Catena, R. Gauchola, and K. Mu, "iRead4Skills - Basic Lexicons per Complexity Level," Mar. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.10889986

[26] A. Pintard, T. François, J. Nagant de Deuxchaisnes, S. Barbosa, M. L. Reis, M. Moutinho, R. Monteiro, R. Amaro, S. Correia, S. Rodríguez Rey, M. Garcia González, K. Mu, and X. Blanco Escoda, "iRead4Skills Dataset 1: Corpora by Complexity Level for FR, PT and SP," Sep. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.13768477

[27] X. Blanco Escoda, R. Amaro, T. François, and M. Garcia, "iRead4Skills - Baselines for complexity lexicons definition," Nov. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.10069793

[28] O. Dahl, "Definitions of complexity," in *Proceedings of the Colloquium on Complexity, Accuracy and Fluency in Second Language Use, Learning & Teaching*, 2007, pp. 41–46.

[29] E. H. Hiebert, "Using multiple sources of information in establishing text complexity," *Reading Research Report*, vol. 11, 2011.

[30] E. H. Hiebert and P. D. Pearson, "Understanding text complexity," *Elementary School Journal*, vol. 115, p. 153–160, 2014.

[31] S. R. Goldman and C. D. Lee, "Text complexity: State of the art and the conundrums it raises," *Elementary School Journal*, vol. 115, p. 290–300, 2014.

[32] M. J. Grosso, A. Soares, F. de Sousa, and J. Pascoal, "QuaREPE - Quadro de Referência para o Ensino de Português no Estrangeiro," MEC/DGIDC Documento Orientador, 2011.

[33] I. Direção de Serviços de Língua e Cultura, Camões, "Referencial Camões Português Língua Estrangeira," Camões Inst. Cooperação e da Língua I.P. Lisboa, 2017.

[34] J. W. Cunningham and H. A. Mesmer, "Quantitative measurement of text difficulty: What's the use?" *Elementary School Journal*, vol. 115, pp. 255–269, 2014.

[35] D. S. McNamara, M. M. Louwerse, P. M. McCarthy, and A. C. Graesser, "Coh-metrix: Capturing linguistic features of cohesion," *Discourse Processes*, vol. 47(4), pp. 292–330, 2010.

[36] M. Flor, B. Klebanov, and K. Sheehan, "Lexical tightness and text complexity," in *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, Atlanta, GA, 2013, pp. 29–38.

[37] X. Chen and D. Meurers, "Word Frequency and Readability: Predicting the Text-level Readability with a Lexical-level Attribute," *Journal of Research in Reading*, vol. 41, no. 3, pp. 486–510, 2018.

[38] Y. Douglas and S. Miller, "Syntactic and lexical complexity of reading correlates with complexity of writing in adults," *International Journal of Business Administration*, vol. 7(4), 2016.

[39] J. M. Imperial, "Knowledge-rich bert embeddings for readability assessment," arXiv preprint arXiv:2106.07935, 2021.

[40] R. Santos, J. Rodrigues, A. Branco, and R. Vaz, "Neural Text Categorization with Transformers for Learning Portuguese as a Second Language," in *EPIA 2021*, 2021, pp. 715–726.

[41] J. Lee and S. Vajjala, "A neural pairwise ranking model for readability assessment," in *Findings of ACL 2022*, 2022, pp. 3802–3813.