The iRead4Skills Text Complexity Analyzer for European Portuguese

Eugénio Ribeiro[1,2], David Antunes[1], Nuno Mamede[1,3], and Jorge Baptista[1,4]

[1] INESC-ID Lisboa, Portugal
[2] Instituto Universitário de Lisboa (ISCTE-IUL), Portugal
[3] Instituto Superior Técnico, Universidade de Lisboa, Portugal
[4] Faculdade de Ciências Humanas e Sociais, Universidade do Algarve, Portugal
{eugenio.ribeiro,david.f.l.antunes,nuno.mamede,jorge.baptista}@inesc-id.pt

Abstract. This demo showcases the iRead4Skills intelligent complexity analyzer, with a focus on the Portuguese language. Aimed at adult native speakers with low literacy and their educators, the tool supports the selection of appropriate reading materials by assessing text complexity across multiple dimensions. It offers analysis at three levels of granularity: an overall complexity classification, a detailed complexity overview, and annotation of difficult linguistic phenomena. By combining language technology with real-world educational needs, the system promotes a more inclusive and data-driven literacy education.

Keywords: Text complexity · Adult literacy · Language technology.

1 Introduction

The iRead4Skills project[5] aims to enhance reading skills among adults by developing an intelligent system that assesses text complexity and recommends suitable reading materials. This contributes to reducing skill gaps and promoting access to information and culture.

In this demo, we will present the intelligent complexity analyzer developed within the project. This tool evaluates textual complexity across multiple dimensions and supports both learners and educators in selecting texts that match the target audience's reading level. While the system is designed for three languages—French, Portuguese, and Spanish—this demonstration will focus on Portuguese.

Figure 1 displays the tool's current web interface. It provides complexity analysis at three levels of granularity: overall classification, a detailed complexity overview, and annotation of difficult linguistic phenomena. These components are described in more detail in the following sections.

2 Complexity Classification

The overall complexity classification distinguishes three levels—very easy, easy, and plain—defined with respect to the target low-literacy population [2], and

[5] https://iread4skills.com/

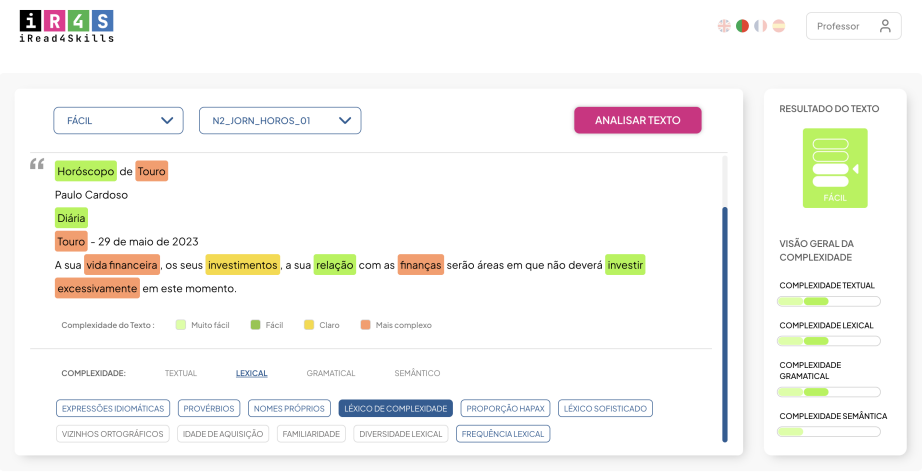


Fig. 1. The iRead4Skills intelligent complexity analyzer.

roughly aligned with A1, A2, and B1 levels of the Common European Framework of Reference for Languages (CEFR) [1]. An additional level is included for texts of higher difficulty. The classification is performed by a hybrid model that combines linguistically motivated features[6] with fine-tuned language models [3].

3 Complexity Overview

The complexity overview offers a more fine-grained analysis of text complexity across four dimensions: textual structure, lexicon, syntax, and semantics. Each dimension relies on a tailored subset of features, selected based on linguistic theory and data-driven insights. Complexity scores for each dimension are determined using a threshold-based approach derived from feature distributions across the different levels.

4 Difficult Phenomena Annotation

The annotation of difficult phenomena is guided by a set of heuristics aligned with the four dimensions used in the complexity overview. These heuristics identify both generally challenging aspects and level-specific difficulties, informed by linguistic resources and threshold-based methods. The interface allows users to select a specific phenomenon and highlight the parts of the text that are difficult with respect to that aspect for the target proficiency level.

[6] https://gitlab.hlt.inesc-id.pt/iread4skills/docs

Acknowledgments. This work was supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT) (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). The interface was designed by Mindshaker and takes the insights of the iRead4Skills consortium and stakeholders into consideration.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Council of Europe: Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press (2001), https://rm.coe.int/1680459f97
2. Monteiro, R., Amaro, R., Correia, S., Pintard, A., Gauchola, R., Moutinho, M., Blanco Escoda, X.: iRead4Skills Complexity Levels. Project Deliverable D3.1, iRead4Skills (2023). https://doi.org/10.5281/zenodo.10459090
3. Ribeiro, E., Mamede, N., Baptista, J.: Avaliação Automática do Nível de Complexidade de Textos em Português Europeu. Linguamática **16**(2), 121–145 (2024). https://doi.org/10.21814/lm.16.2.449