PORTHO: A Corpus-Based Resource of Orthographic Neighbors in 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

Orthographic neighbors (ONs) play a central role in models of visual word recognition and have been shown to influence reading speed, lexical access, and literacy development. Despite their importance, resources providing detailed and flexible ON information remain scarce for European Portuguese. This paper introduces PORTHO, a corpus-based lexical resource that provides multiple ON metrics for over 43,000 word forms, using several ON definitions. In addition to classical neighborhood size measures, PORTHO provides frequency-based statistics and graded orthographic distance (OD) features. We analyze the statistical properties of the resource and evaluate its empirical utility in automatic text complexity assessment using the iRead4Skills[1] corpus. Results show that while ON features alone are insufficient to predict readability, they contribute complementary information and compare favorably with existing resources for Portuguese. PORTHO is made publicly available in different formats to support research in psycholinguistics, readability modeling, and Natural Language Processing (NLP) for Portuguese.

1 Introduction

Orthographic neighbors (ONs) were originally introduced by Coltheart et al. (1977) as words of the same length that differ from a target word by a single letter substitution (e.g., *hat* and *hot*). The number of such neighbors, commonly referred to as Coltheart's N, has been shown to influence word recognition speed and accuracy in a wide range of psycholinguistic tasks. From a computational perspective, this definition corresponds to a constrained instance of the Levenshtein distance (LD) (Levenshtein, 1966), where only substitutions are allowed and word length is held constant. Subsequent work has generalized this notion by allowing insertions and deletions and by adopting graded similarity measures, such as Orthographic LD 20 (OLD20) (Yarkoni et al., 2008), defined as the average edit distance between a word and its closest orthographic neighbors.

ONs play a central role in models of visual word recognition and lexical access. Overlapping orthographic representations may activate multiple lexical candidates in parallel, leading to competition and facilitation effects during word identification (Grainger, 2008; Norris, 2013). Empirical studies have shown that words embedded in dense orthographic neighborhoods are often recognized more quickly and accurately, particularly when they are short and frequent, highlighting the interaction between neighborhood structure and other lexical properties (Tulkens et al., 2020; Grainger, 2024). These effects have made ON measures standard control variables in psycholinguistic experiments and relevant indicators in broader research on reading acquisition and literacy.

Despite their importance, computational resources providing ON information remain unevenly distributed across languages. While several large-scale and flexible resources exist for English and other widely studied languages, resources for European Portuguese are comparatively limited. Existing databases—PORLEX (Gomes and Castro, 2003) and P-PAL (Soares et al., 2018)—either focus on restricted neighborhood definitions or are difficult to access programmatically, limiting their usefulness for large-scale computational studies and Natural Language Processing (NLP) applications.

To address this gap, this paper introduces PORTHO, a large-scale, corpus-based ON resource for European Portuguese. PORTHO integrates multiple neighborhood definitions and orthographic distance (OD) metrics derived from Levenshtein and Damerau-Levenshtein (Damerau, 1964) distances and provides a rich set of frequency-based and distance-based features for over 43,000 word forms. In addition to making the resource pub-

[1] https://iread4skills.com/

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 406–415

licly available[2], we assess its empirical utility in automatic text complexity classification using the iRead4Skills corpus (Pintard et al., 2024; Baptista et al., 2024), thereby assessing the empirical utility of PORTHO for readability-oriented tasks.

The remainder of the paper is structured as follows. Section 2 reviews related work on ON resources and their use in readability and text complexity research. Section 3 describes the construction of the PORTHO resource, while Section 4 presents an analysis of its main statistical properties. Section 5 examines the relationship between ON features and text complexity through correlation analysis and automatic classification experiments. Section 6 concludes the paper and outlines directions for future work.

2 Related Work

In this section, we provide an overview of existing lexical resources that include ON information, with particular focus on Portuguese, in Section 2.1. Then, in Section 2.2, we discuss how such information has been exploited in readability and text complexity research.

2.1 Orthographic Neighbor Resources

Similarly to most linguistic subjects, the available ON resources are predominantly in English. For instance, MCWORD (Medler and Binder, 2005) provides Coltheart's N information for 66,372 word forms; N-WATCH (Davis, 2005) provides 12 neighborhood statistics for 30,605 words, including Coltheart's N, the number of neighbors by swapping the letter in each position, and frequency-based filters, and provides tools for computing the same statistics for new vocabularies; and IPHOD (Vaden et al., 2009) provides phonological neighborhood density information for 54,030 words.

Moving to other languages, CLEARPOND (Marian et al., 2012) is an online database that provides orthographic and phonological neighborhood information within and between five languages: Dutch, English, French, German, and Spanish. It defines ONs as words at an LD of 1 and provides lists of neighbors, neighborhood densities, and mean neighborhood frequencies for 27,751 words per language. It also allows filtering by LD operation (substitution, addition, and deletion), frequency, and lexical information.

Considering individual languages, LEXIQUE3 (New et al., 2005) provides Coltheart's N, the OLD20, and the Orthographic Uniqueness Point (OUP) (the letter position in a word where it becomes distinct from all other words in the vocabulary), as well as their phonological counterparts, for 140,000 French words. ONESC (Martín and Pérez, 2008) provides ON information for approximately 100,000 Spanish words typically read by children. It provides Coltheart's N, as well as filters based on the positions of divergent letters, the cumulative frequency of the neighbors, and the number of neighbors of higher, lower, and equal frequency.

Focusing on Portuguese, the Brazilian Portuguese Lexicon (LEXPORBR) (Estivalet and Meunier, 2016) provides Coltheart's N, the OLD20, and the OUP for 215,175 word types. For the European variety of the language, there are two main resources: PORLEX (Gomes and Castro, 2003) and Procura-PALavras (P-PAL) (Soares et al., 2018).

PORLEX provides lists of ONs, defined as words at an LD of 1, as well as the OUP, for 27,374 words. Additionally, it provides both phonetic and phonemic counterparts of these features. Still, it is limited in terms of the ON features it provides.

P-PAL is a web interface to a database that covers 208,642 word forms and provides several neighborhood features, such as Coltheart's N, the OLD20, the OUP, frequency-based filters, and the spread (the number of letters in the word which generate valid neighbors when replaced). However, it is currently unavailable and it was not designed for direct computational applications.

The limitations of the existing resources open the door for a new ON resource that is more flexible, both in terms of the features it provides and how it can be accessed.

2.2 Orthographic Neighbors and Readability

The notion of ONs has gradually been incorporated into computational approaches to readability and text complexity assessment (North et al., 2023). While much of the early work focused on English, subsequent studies have shown that the role of ONs varies substantially across languages and writing systems. This variation challenges the assumption that ON effects are universal and highlights the need for language-specific resources and analyses.

Differences in orthographic depth, morphological richness, and frequency distributions shape how orthographic similarity influences reading. Frost (1992) argues that orthographic systems should be

[2] https://string.hlt.inesc-id.pt/demo/portho

viewed not merely as transparent or opaque, but as part of a broader cognitive ecology. In this view, neighborhood density may interact with morphological structure and visual complexity in ways that differ across languages. Indeed, while dense neighborhoods are often associated with facilitation effects in alphabetic languages, evidence from other writing systems complicates this picture. For example, in Korean—a script combining alphabetic and syllabic properties—higher neighborhood density has been linked to slower reading performance, suggesting increased cognitive load rather than facilitation (Tulkens et al., 2020). Such findings suggest that the effects of orthographic similarity depend on how visual and morphological information is encoded and processed (Schmalz et al., 2024).

A better understanding of ON effects across languages could inform the development of literacy assessments and reading interventions tailored to specific linguistic contexts (Cox and Probert, 2024). In particular, computational tools that incorporate orthographic features (e.g., neighborhood density (Biedermann et al., 2024), word length, and frequency) could support adaptive reading technologies and multilingual education platforms.

In the context of automatic readability assessment, ONs are typically considered alongside other lexical indicators such as word length, frequency, and morphological complexity. For instance, ON features aggregated at the document level are leveraged in the FABRA readability toolkit for French (Wilkens et al., 2022) and in the iRead4Skills Intelligent Complexity Analyzer for French, Portuguese, and Spanish (Aissa et al., 2025).

However, evidence regarding the independent contribution of orthographic neighborhood measures to readability remains mixed. Some studies report that their explanatory power diminishes once semantic or morphological variables are taken into account, at least for skilled adult readers (Pereira and Justi, 2024). In contrast, research on children's reading development, particularly in transparent and semi-transparent orthographies, such as Portuguese, points to substantial effects on early reading and spelling, underscoring the usefulness of ON metrics in assessing emergent literacy and reading difficulty (Fernandes et al., 2008; Querido et al., 2020, 2021).

From a computational perspective, these mixed results underscore the importance of flexible resources that support different neighborhood definitions and aggregation strategies. Most existing studies rely on a limited set of orthographic measures, which restricts systematic comparison across metrics and languages. By providing multiple ON and OD features for European Portuguese and evaluating their behavior in a readability classification task, the present work contributes to a more nuanced understanding of how orthographic similarity relates to text complexity.

3 Resource Construction

This section outlines the systematic approach adopted to develop the lexical resource for ONs in European Portuguese.

At the base of the resource are the lexical frequencies observed in the set of European Portuguese corpora compiled by Linguateca[3] in the context of project AC/DC (Santos and Sarmento, 2002). These lexical frequencies were computed for over one million word forms. However, not all these are relevant for building an ON resource. Thus, we applied a filtering process based on the information provided by the STRING NLP Chain (Mamede et al., 2012). We removed out-of-vocabulary words (mostly spelling errors), symbols and punctuation, digits and roman numerals, acronyms and abbreviations, proper and compound nouns, and single-character tokens[4]. Finally, we discarded surface forms with lexical frequency lower than 100, ending up with a total of 43,429 distinct forms.

To identify the ONs, we started by computing the pairwise edit distance between the surface forms using the implementation provided by the Natural Language Toolkit (NLTK)[5] (Bird and Loper, 2004). We used two different edit distances: Levenshtein (Levenshtein, 1966) and Damerau-Levenshtein (Damerau, 1964). While the former only considers insertions, deletions, and substitutions, the latter also considers transpositions. These distances were used to compute three increasingly less restrictive sets of orthographic neighbors for each surface form:

- The orthographic neighbors according to Coltheart's definition, i.e., words with the same length at an LD of 1 (Coltheart et al., 1977);
- The Levenshtein neighbors, i.e., words at an LD of 1;

[3] https://www.linguateca.pt/

[4] The surface forms were only filtered out if all of their analyses belong to the discard categories.

[5] https://www.nltk.org/

- The Damerau-Levenshtein neighbors, i.e., words at a Damerau-Levenshtein distance (DLD) of 1.

These three sets were then used to compute the features described in Table 1 for each surface form.

Feature	Description
ON	Number of ONs
ON AvgFreq	Average frequency of the ONs
ON CumFreq	Cumulative frequency of the ONs
>ON	Number of ONs more frequent than the word
>ON AvgFreq	Average frequency of more frequent ONs
>ON CumFreq	Cumulative frequency of more frequent ONs

Table 1: Base ON features provided by PORTHO.

Additionally, for each surface form and distance metric, we obtained the set of 100 nearest neighbors and the corresponding distance value, with ties ordered by increasing word length difference. This allowed us to compute orthographic similarity features such as the OLD20 proposed by Yarkoni et al. (2008), that is, the average LD to the 20 nearest neighbors. More specifically, for each surface form and distance metric, we computed the average OD to the n nearest neighbors for $n \in \{10, 20, 50, 100\}$. Even though Yarkoni et al. (2008) did not observe significant differences in explained variance for $n \in [5, 50]$, they only explored lexical decision and pronunciation performance tasks. Thus, we opted for computing the OD features for different values of n, allowing potential users of the resource to select the most appropriate for their tasks.

4 Resource Analysis

As discussed in the previous section, PORTHO provides ON features for 43,429 distinct forms. Table 2 shows statistics of the multiple features in the resource, namely the average, minimum, and maximum values. We can see that, as expected, using Coltheart's definition produces less ONs, as it only considers words of the same length. On the other hand, from this high level, there seem to be no significant differences between using the Levenshtein or Damerau-Levenshtein distances.

When considering more frequent ONs only, we can see that while on average the number halves, the maximum number only decreases by 20%. Consistently, their average frequency doubles. On the other hand, their cumulative frequency only decreases by 10% on average and there is no significant difference in terms of maximum values.

Looking into OD features, there are no significant differences between the two distance metrics, with only a residual decrease in terms of the average values when using the DLD. Regarding minimum and maximum values, the only change is in the maximum value when considering 100 neighbors. This, together with the statistics of the ON features, suggests that considering transpositions has only minor impact in terms of overall ONs.

5 Relation with Readability

To assess the relation between the information regarding ONs provided by PORTHO and readability or text complexity, we rely on the iRead4Skills corpus (Pintard et al., 2024) and perform two analyses: one based on correlation, and the other on classification performance.

5.1 Dataset

The iRead4Skills corpus (Pintard et al., 2024), collected and annotated in the context of the project with the same name[6] (Baptista et al., 2024), consists of a large collection of texts in three languages—French, Portuguese, and Spanish. The texts were classified by human experts into four complexity levels, roughly corresponding to Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) levels, but targeted at adult native speakers with low literacy (Monteiro et al., 2023):

Very Easy: Texts that are fully or almost fully understood by everyone, including people with very low schooling (i.e., that did not finish the primary school) and almost no reading experience. It corresponds to CEFR A1 level.

Easy: Texts that are fully or almost fully understood by people with low schooling (i.e., that completed the primary school but do not have more than the 9th year) and have poor reading experience. It corresponds to CEFR A2 level.

Plain: Texts that are understood the first time they are read by people that completed the 9th year and have a functional-to-average reading experience. It corresponds to CEFR B1 level.

More Complex: Texts with a higher complexity than that defined by the previous levels. It corresponds to CEFR levels B2 and above.

[6] https://iread4skills.com/

Feature	Coltheart	Levenshtein	Damerau-Lev.
ON	1.28 (0–25)	2.27 (0–36)	2.29 (0–36)
ON AvgFreq	3,407.16 (0–5,784,687)	4,553.40 (0–5,784,687)	4,516.57 (0–5,784,687)
ON CumFreq	19,710.08 (0–18,352,798)	33,724.82 (0–18,353,297)	33,952.59 (0–18,353,297)
>ON	0.64 (0–20)	1.14 (0–29)	1.14 (0–29)
>ON AvgFreq	6,591.27 (0–10,277,101)	9,574.39 (0–10,277,101)	9,503.66 (0–10,277,101)
>ON CumFreq	18,046.75 (0–18,297,657)	30,986.73 (0–18,297,657)	31,198.34 (0–18,297,657)
OD10		2.30 (1–9)	2.29 (1–9)
OD20		2.68 (1–10)	2.68 (1–10)
OD50		3.19 (1–10)	3.19 (1–10)
OD100		3.58 (2–11)	3.56 (2–10)

Table 2: Statistics—*Average (Minimum–Maximum)*—of the multiple ON features provided by PORTHO.

	V. Easy	Easy	Plain	+Complex	Total
Train	523	418	490	555	1986
Val.	142	111	123	143	519
Test	112	137	139	40	428

Table 3: Distribution of the Portuguese texts in the iRead4Skills dataset by partition and complexity level.

Focusing on Portuguese data, the iRead4Skills dataset contains 2,933 examples (Reis et al., 2024). In previous studies (Ribeiro et al., 2025), this dataset was split into three subsets, distributed as shown in Table 3. We can see that the distribution across complexity levels is fairly balanced in the train and validation sets, but there are fewer examples of the more complex level in the test set.

5.2 Feature Aggregation

Portho provides ON features at the word level. However, the iRead4Skills corpus is annotated for complexity at the document level. Thus, in order to assess the relation between the two, the word-level features need to be aggregated. In previous studies on readability assessment using the iRead4Skills corpus (Ribeiro et al., 2025), word-level features were aggregated using 17 different aggregators covering distributional aspects, such as range, separation, central tendency, dispersion, and curve description. For simplicity, in this study, we rely on just 3 aggregators: sum, maximum, and average.

5.3 Correlation Analysis

To assess the correlation between the ON features and textual complexity, we rely on the Kendall rank correlation coefficient (Kendall, 1938), specifically the τ_B statistic (Kendall, 1945). We opted for this statistic because it is more robust to the several ties introduced by the reduced number of complexity levels than the more commonly used Pearson and Spearman coefficients. Table 4 shows the correlation between the ON features computed using the three sets of ONs provided by PORTHO and the complexity levels defined in the context of the iRead4Skills project. Additionally, it shows the correlation of the same features computed using the ON set provided by PORLEX.

First of all, we can see that the aggregators play a dominant role in comparison to the base features, with sum-aggregated features showing the highest—although still moderate—correlation with textual complexity. We argue that this is due to the strong correlation between sum-aggregated features and textual length, which has long been established as an effective predictor of textual complexity (Flesch, 1948). While the correlation between the sum-aggregated number of ONs and textual length in number of words is above 0.85 for every set of ONs, the correlation between textual length and textual complexity is 0.43 on the iRead4Skills corpus, which is in line with that of the sum-aggregated features with highest correlation.

When the average or maximum aggregators are used, the correlation is weak at most. This suggests that, although ONs can contribute to textual complexity, their relevance in this context is less prominent than that of other indicators, such as textual length.

Comparing the features computed from the PORTHO and PORLEX ON sets, we can observe similar correlations for the most correlated features, but, as we progress toward the least correlated ones, the decrease is more pronounced for the features

Feature	Coltheart	Levenshtein	Damerau-Lev.	Porlex
ON CumFreq Sum	0.44	0.44	0.43	0.44
ON AvgFreq Sum	0.44	0.42	0.42	0.45
ON Sum	0.38	0.40	0.40	0.43
>ON AvgFreq Sum	0.30	0.32	0.32	0.41
>ON CumFreq Sum	0.27	0.28	0.28	0.41
>ON Sum	0.24	0.28	0.28	0.33
>ON AvgFreq Max	0.21	0.21	0.22	0.23
ON CumFreq Avg	0.19	0.17	0.16	0.19
ON Max	0.18	0.19	0.19	0.12
ON AvgFreq Avg	0.16	0.12	0.11	0.22
ON AvgFreq Max	0.11	0.13	0.13	0.19
ON CumFreq Max	0.11	0.16	0.11	0.18
>ON CumFreq Max	0.10	0.10	0.10	0.16
>ON Max	0.08	0.09	0.09	0.10
>ON AvgFreq Avg	-0.02	-0.03	-0.03	0.08
ON Avg	-0.14	-0.07	-0.08	0.16
>ON CumFreq Avg	-0.05	-0.08	-0.08	0.07
>ON Avg	-0.21	-0.22	-0.22	-0.14

Table 4: Correlation (Kendall's τ) between the ON features and textual complexity. The features are presented in descending order of the average correlation across the four ON sets.

computed from the PORTHO sets. Although PORLEX considers a smaller vocabulary, it considers single-character words, which have a large amount of ONs and include very common determiners (e.g., *o* and *a*, 'the') and conjunctions (e.g., *e*, 'and'). The less pronounced decrease in correlation suggests that these single-word characters may have some relation with readability, possibly because an increased amount of determiners and conjunctions is an indicator of more complex or, at least, longer documents. However, as the inclusion or exclusion of single-word characters is not the only difference between the PORTHO and PORLEX ON sets, further analysis is required to confirm this possibility.

Overall, the features that only consider ONs with higher frequency reveal lower correlation than their counterparts that consider every ON. The only exception is the maximum average frequency, which has the highest correlation among the features that are not sum-aggregated. We also found a significant negative correlation between the average orthographic neighbourhood size and our measure of difficulty. This suggests that words with more ONs tend to be easier to process, which is consistent with psycholinguistic models of lexical access where higher neighbour density can facilitate word recognition by increasing co-activation of similar lexical representations (Lim, 2016).

Feature	Levenshtein	Damerau-Lev.
OD10 Sum	0.43	0.43
OD100 Sum	0.43	0.43
OD50 Sum	0.43	0.43
OD20 Sum	0.43	0.43
OD100 Max	0.36	0.36
OD50 Max	0.35	0.35
OD20 Max	0.35	0.35
OD10 Max	0.35	0.35
OD50 Avg	0.25	0.25
OD100 Avg	0.24	0.24
OD10 Avg	0.24	0.24
OD20 Avg	0.23	0.23

Table 5: Correlation (Kendall's τ) between the OD features and textual complexity. The features are presented in descending order of the average correlation across the two distance metrics.

Moving to the OD features, Table 5 shows no observable difference between the two distance metrics in terms of correlation with the complexity level. As discussed in Section 4, considering transpositions has only minor impact on the OD features, which justifies these results.

Similarly to what was observed by (Yarkoni et al., 2008) for the lexical decision and pronun-

ciation performance tasks, the number of neighbors has only minimal impact on correlation. This suggests that the distance to the closest neighbors contributes the most, while additional neighbors are only relevant to capture subtle differences.

Finally, similarly to what was observed for the ON features, the used aggregator has the most impact on correlation. Using the sum aggregator leads to the highest correlation, in line with that of textual length. In fact, the correlation between the sum-aggregated OD features and textual length is 0.92, which is higher than that observed for the ON features. When using the maximum and average aggregators, the correlation is higher than that observed for the ON features and, in this case, the advantage of the maximum aggregator is clear, achieving moderate correlation.

Overall, an higher OD means that the word has less ON, which is more common for longer words. Thus, similarly to the negative correlation observed for the average number of ONs, the positive correlation observed between OD features and textual complexity suggests that any potential confusion introduced by words with several ONs is outweighed by the complexity introduced by longer words. In Portuguese, these are typically related to more morphologically complex formations, such as derivational and inflectional processes that increase semantic density and syntactic specificity. Furthermore, the higher correlation observed when using the maximum aggregator suggests that one single difficult word has a greater impact on the perceived complexity than the overall complexity of the words in the text.

5.4 Automatic Classification

Although ON features have been used for textual complexity assessment in several languages (e.g., Wilkens et al., 2022; Ribeiro et al., 2024, 2025), their ability to predict complexity independently has not been assessed. Thus, we trained random forest classifiers on the iRead4Skills corpus using the features generated using each ON set in PORTHO, as well as the OD features and combinations of both. These are compared with classifiers trained on both the ON features generated using the PORLEX neighbor set and the full set of 631 (199 base features, 27 of which are aggregated using 17 different aggregators) descriptive, lexical, syntactic, and discursive features identified as potential complexity indicators in the context of the iRead4Skills project (Ribeiro et al., 2025).

We adopt some of the most common evaluation metrics across previous studies on automatic readability level classification, namely accuracy, adjacent accuracy, and the macro F_1 score. Accuracy evaluates the precise identification of a text's complexity level, while adjacent accuracy also considers neighboring levels, offering further insight into the identification of texts slightly easier or harder than the assigned level. As the distribution of examples across levels is not balanced in the test set, the macro F_1 score is also a relevant metric to identify potential biases. Considering the non-deterministic nature of random forest classifiers, for each configuration, we trained 20 different classifiers. Table 6 shows the average and standard deviation values of each metric across the 20 runs on both the validation and test sets.

Comparing the results in the first block of Table 6, we can see that there are no significant differences between the results achieved using the different neighbor sets provided by PORTHO. However, they outperform the results achieved using the PORLEX neighbor set, especially on the validation set. Considering that, in Section 5.3, a higher correlation with textual complexity was observed when using the PORLEX neighbor set, these results suggest that there are nuanced interactions between ON features and textual complexity, in a relation that is not always linear. The neighbor sets provided by PORTHO seem to capture these nuances to a slightly greater extent than the PORLEX set.

Moving to the OD features, in the second block of Table 6, we can see that using the LD leads to slightly better performance than the DLD. Additionally, even though the performance is similar to that of the ON features on the validation set, the classifiers trained on the OD features generalize better to the test set. This happens because the ON features cover several aspects regarding orthographic neighboring relations, while the OD features summarize that information into a single, more generic aspect and, thus, are less prone to lead to overfitting.

In the third block of Table 6, we can see that, considering a given distance function, the combination of ON and OD features leads to improved performance on the validation set, suggesting that they are, in fact, complementary. However, that comes at the expense of generalization ability, leading to a symmetric decrease in performance on the test set. Perhaps surprisingly considering the previously discussed results, the results in the fourth

Feature Set	Validation Set			Test Set		
	Acc	Adj Acc	F_1	Acc	Adj Acc	F_1
Coltheart	45.15±0.56	82.16±0.42	42.98±0.64	41.48±0.63	84.08±0.44	40.14±0.61
Levenshtein	45.48±0.61	80.94±0.42	43.39±0.64	41.31±0.58	84.66±0.29	39.98±0.55
Damerau-Lev.	45.57±0.73	81.14±0.37	43.54±0.69	40.46±0.53	84.80±0.28	39.15±0.49
OLD	45.75±0.53	82.21±0.30	44.10±0.57	43.08±0.86	84.60±0.44	42.22±0.88
ODLD	45.39±0.42	82.04±0.27	43.58±0.46	42.30±0.41	84.81±0.43	41.42±0.43
Levenshtein + OLD	47.57±0.61	83.77±0.30	45.74±0.64	40.91±0.48	86.54±0.48	39.74±0.47
Damerau-Lev. + ODLD	47.95±0.62	83.54±0.30	45.97±0.68	40.99±0.57	86.52±0.37	39.68±0.62
All	49.97±0.15	83.20±0.05	48.18±0.19	43.08±0.44	86.39±0.21	41.98±0.38
PORLEX	43.12±0.76	81.46±0.54	40.88±0.72	40.06±0.75	84.28±0.48	38.54±0.76
iR4S	54.53±0.54	87.41±0.45	52.11±0.60	49.53±1.51	91.59±0.38	48.06±1.73

Table 6: Text complexity classification results. The presented results correspond to the average and standard deviation values across 20 runs.

block of Table 6 reveal that the combination of every feature, that is, the ON features computed using the three neighbor sets and the OD features computed using the two distance functions, leads to the best performance, in spite of the ever present signs of overfitting to a certain extent. Furthermore, it leads to the most stable results, with the lowest standard deviation across runs. These results suggest that, even though many of the features seem to provide similar information, they are actually complementary and can help distinguishing the nuances of textual complexity.

Overall, a performance below 50% in terms of accuracy and F_1 shows that information regarding ONs alone is not enough to predict textual complexity, which is a complex and nuanced problem. Still, several studies (e.g., Branco et al., 2014; Ribeiro et al., 2024) have shown that readability or textual complexity assessment is a very difficult and subjective task. This is confirmed by the performance achieved using the full iRead4Skills feature set, which, on average, is only 5 percentage points better than that achieved using the full set of ON features. This suggests that the computed features are able to provide information that goes beyond simple neighboring relations. For instance, as previously discussed, the sum-aggregated features are proxies for textual length, which is a relevant factor for textual complexity.

6 Conclusion and Future Work

This paper presented PORTHO, a large-scale, corpus-based ON resource for European Portuguese that integrates multiple neighborhood metrics derived from several edit-distance variants. It is made publicly available in different formats to support research in psycholinguistics, readability modeling, and NLP for Portuguese.

Additionally, we evaluated the empirical utility of the resource for automatic text complexity assessment using the iRead4Skills corpus. Our analyses indicate that, while ONs alone are not sufficient predictors of textual complexity, the information encoded in PORTHO contributes meaningfully to models of readability and compares favorably with existing resources such as PORLEX.

As future work, we intend to extend PORTHO with additional ON related metrics, such as the OUP and the spread. Moreover, we intend to build two complementary resources: one for ONs of lemmas instead of surface forms, and one for phonological neighbors. Then, we intend to assess whether these extensions can provide complementary information for readability assessment purposes and other related NLP tasks.

Limitations

A few limitations of the current version of PORTHO should be acknowledged. Because all neighborhood relations are computed automatically, the resource may inadvertently include edge cases arising from orthographic variation, noise in lexical frequency data, or segmentation inconsistencies. Moreover, foreign words and borrowings present in the source corpora can introduce neighborhood patterns that do not reflect the morphological or phonological structure of European Portuguese. Finally, some marginal lexical categories—such as interjec-

tions—pose challenges for neighborhood computation, as their forms often lack stable paradigmatic relations. These considerations motivate some of the refinements outlined in the future work plan.

Acknowledgments

This work was supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT) under projects UID/50021/2025 (DOI: 10.54499/UID/50021/2025) and UID/PRR/50021/2025 (DOI: 10.54499/UID/PRR/50021/2025) and by the European Commission (Project: iRead4Skills, Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837).

References

Wafa Aissa, Raquel Amaro, David Antunes, Thibault Bañeras-Roux, Jorge Baptista, Alejandro Catala, Luís Correia, Thomas François, Marcos Garcia, Mario Izquierdo-Álvarez, Nuno Mamede, Vasco Martins, Miguel Neves, Eugénio Ribeiro, Sandra Rodriguez, and Elodie Vanzereven. 2025. The iRead4Skills Intelligent Complexity Analyzer. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 73–84.

Jorge Baptista, Eugénio Ribeiro, and Nuno Mamede. 2024. iRead4Skills @ IberSPEECH 2024: Project Presentation and Developments for the Portuguese Language. In *Proceedings of the IberSPEECH Conference*, pages 297–299.

Britta Biedermann, Elke Beyersmann, Madara Blosfelds, Carla Macapagal, Alexandra Rosevear, and Willow Marinovic. 2024. Cross-language Orthographic Neighborhood Density Effects in Dutch–English and Spanish–English Bilinguals. *Frontiers in Language Sciences*, 3.

Steven Bird and Edward Loper. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217.

António Branco, João Rodrigues, Francisco Costa, João Silva, and Rui Vaz. 2014. Assessing Automatic Text Classification for Interactive Language Learning. In *Proceedings of the International Conference on Information Society (i-Society)*, pages 70–78.

Max Coltheart, Eileen Davelaar, Jon Torfi Jonasson, and Derek Besner. 1977. *Access to the Internal Lexicon*, pages 535–555. Routledge.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

Penelope Cox and Tessa Probert. 2024. Expanding the Neighbourhood Watch: Orthographic Neighbours in isiXhosa Reading and Spelling. *Reading and Writing*, 15(1):a461.

Fred J. Damerau. 1964. A Technique for Computer Detection and Correction of Spelling Errors. *Communications of the ACM*, 7(3):171–176.

Colin J. Davis. 2005. N-Watch: A Program for Deriving Neighborhood Size and Other Psycholinguistic Statistics. *Behavior Research Methods*, 37(1):65–70.

Gustavo L. Estivalet and Fanny Meunier. 2016. The Brazilian Portuguese Lexicon: An Instrument for Psycholinguistic Research. *PLOS ONE*, 10(12):e0144016.

Sandra Fernandes, Paulo Ventura, Luís Querido, and José Morais. 2008. Reading and Spelling Acquisition in European Portuguese: A Preliminary Study. *Reading and Writing*, 21(8):805–821.

Rudolph Flesch. 1948. A New Readability Yardstick. *Journal of Applied Psychology*, 32(3):221—-233.

Ram Frost. 1992. *Orthography and Phonology: The Psychological Reality of Orthographic Depth*, pages 255–274. John Benjamins Publishing Company.

Inês Gomes and São Luís Castro. 2003. Porlex, a Lexical Database in European Portuguese. *Psychologica*, 32:91–108.

Jonathan Grainger. 2008. Cracking the orthographic code: An introduction. *Language and Cognitive Processes*, 23(1):1–35.

Jonathan Grainger. 2024. Letters, Words, Sentences, and Reading. *Journal of Cognition*, 7(1):1–21.

M. G. Kendall. 1938. A New Measure of Rank Correlation. *Biometrika*, 30(1/2):81–93.

M. G. Kendall. 1945. The Treatment of Ties in Ranking Problems. *Biometrika*, 33(3):239–251.

Vladimir I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, 10(8):707–710.

Stephen Wee Hun Lim. 2016. The Influence of Orthographic Neighborhood Density and Word Frequency on Visual Word Recognition: Insights from RT Distributional Analyses. *Frontiers in Psychology*, 7:401.

Nuno Mamede, Jorge Baptista, Cláudio Diniz, and Vera Cabarrão. 2012. STRING - A Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. In *Proceedings of the International Conference on Computational Processing of Portuguese (PROPOR) - Demo Session*.

Viorica Marian, James Bartolotti, Sarah Chabal, and Anthony Shook. 2012. CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. *PLOS ONE*, 7(8):e43230.

Jesús A. Martínez Martín and M. Emma García Pérez. 2008. ONESC: A Database of Orthographic Neighbors for Spanish Read by Children. *Behavior Research Methods*, 40(1):191–197.

David A. Medler and Jeffrey R. Binder. 2005. MCWord: An On-Line Orthographic Database of the English Language.

Ricardo Monteiro, Raquel Amaro, Susana Correia, Alice Pintard, Roser Gauchola, Michell Moutinho, and Xavier Blanco Escoda. 2023. iRead4Skills Complexity Levels. Project Deliverable D3.1, iRead4Skills.

Boris New, Christophe Pallier, and Ludovic Ferrand. 2005. Manuel de Lexique 3. Technical report, Laboratoire de Psychologie Expérimentale, Université René-Descartes.

Dennis Norris. 2013. Models of Visual Word Recognition. *Trends in Cognitive Sciences*, 17(10):517–524.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical Complexity Prediction: An Overview. *ACM Computing Surveys*, 55(9):1–42.

Rodrigo P. Pereira and Flávia R. R. Justi. 2024. Morphological Priming and Semantic Transparency in Visual Word Recognition in Brazilian Portuguese. *Psicologia: Reflexão e Crítica*, 37:5.

Alice Pintard, Thomas François, Justine Nagant de Deuxchaisnes, Sílvia Barbosa, Maria Leonor Reis, Michell Moutinho, Ricardo Monteiro, Raquel Amaro, Susana Correia, Sandra Rodríguez Rey, Marcos Garcia González, Keran Mu, and Xavier Blanco Escoda. 2024. iRead4Skills Dataset 1: Corpora by Complexity Level for FR, PT and SP. Project Deliverable D3.2, iRead4Skills.

Luís Querido, Sara Fernandes, and Arlette Verhaeghe. 2021. Orthographic Knowledge, and Reading and Spelling: A Longitudinal Study in an Intermediate Depth Orthography. *Spanish Journal of Psychology*, 24:e3.

Luís Querido, Sara Fernandes, Arlette Verhaeghe, and Carlos Marques. 2020. Lexical and Sublexical Orthographic Knowledge: Relationships in an Orthography of Intermediate Depth. *Reading and Writing*, 33(10):2459–2479.

Maria Leonor Reis, Sílvia Barbosa, Michell Moutinho, Ricardo Monteiro, Susana Correia, and Raquel Amaro. 2024. Intelligent Support for Low Literacy Adults: The European Portuguese iRead4Skills Corpus. *International Journal of Emerging Technologies in Learning (iJET)*, 19(8):61—81.

Eugénio Ribeiro, David Antunes, Nuno Mamede, and Jorge Baptista. 2025. Exploring Few-Shot Approaches to Automatic Text Complexity Assessment in European Portuguese. *Journal of the Brazilian Computer Society*, 31(1):690–710.

Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. 2024. Avaliação Automática do Nível de Complexidade de Textos em Português Europeu. *Linguamática*, 16(2):121–145.

Diana Santos and Luís Sarmento. 2002. O Projecto AC/DC: Acesso a Corpora/Disponibilização de Corpora. In *Actas do Encontro Nacional da Associação Portuguesa de Linguística (APL)*, pages 705–717. Associação Portuguesa de Linguística.

Xenia Schmalz, Jay G. Rueckl, and Noam Siegelman. 2024. How We Should Measure Orthographic Depth: Or Should We? OSF Preprints.

Ana Paula Soares, Joana Machado, Ana Costa, Ángel Iriarte, Alberto Simões, José João de Almeida, Manuel Comesaña, and Montserrat Perea. 2018. Procura-PALavras (P-PAL): A Web-based Interface for a New European Portuguese Lexical Database. *Behavior Research Methods*, 50:1461–1481.

Stephan Tulkens, Dirk Sandra, and Walter Daelemans. 2020. Orthographic Codes and the Neighborhood Effect: Lessons from Information Theory. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 172–181.

Kenneth I. Vaden, Harry R. Halpin, and Gregory S. Hickok. 2009. Irvine Phonotactic Online Dictionary.

Rodrigo Wilkens, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin Yancey, and Thomas François. 2022. FABRA: French Aggregator-Based Readability Assessment Toolkit. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1217–1233.

Tal Yarkoni, David Balota, and Melvin Yap. 2008. Moving beyond Coltheart's N: A new Measure of Orthographic Similarity. *Psychonomic Bulletin & Review*, 15(5):971–979.