Automatic readability classifier for European Portuguese

Pedro Curto[1,2], Nuno Mamede[1,2] and Jorge Baptista[1,3]

[1] Universidade de Lisboa, Instituto Superior Técnico
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
Nuno.Mamede@tecnico.ulisboa.pt
[2] INESC-ID Lisboa/L2F – Spoken Language Lab
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
pkurto@gmail.com
[3] Universidade do Algarve/FCHS and CECL
Campus de Gambelas, 8005-139 Faro, Portugal
jbaptis@ualg.pt

Abstract. This paper describes a system that automatically classifies text readability for European Portuguese, while highlighting the key challenges on language features' selection and text classification. To this goal, the system uses existing Natural Language Processing (NLP) tools to extract linguistic features from texts, which are then used by an automatic readability classifier. Currently, the system extracts 52 features grouped in 7 groups: parts-of-speech (POS), syllables, words, chunks and phrases, averages and frequencies, and some extra features. A classifier was created using these features and a corpus, previously annotated by readability level, using a five-level language classification official standard for Portuguese as Second Language. In a five-level (from A1 to C1) and three level (A, B and C) scenarios, the best-performing learning algorithm (LogitBoost) yields 79.25% and 86.32%, respectively.

Keywords: readability assessment metrics, linguistic features extraction, classification, readability, Portuguese

1 Introduction

Readability, or "text difficulty", remains today a relevant research topic, with strong pedagogic impact, especially connected to the development of materials to assist language learning. Studies in this area seek to create a difficulty scale for the assessment of the language level used in texts, since giving students reading materials that are "too difficult" or "too easy" can both hinder the learning process and demotivate the students [11].

According to some studies [16], text readability is affected both by lexical difficulty, related to word difficulty, and by the syntactic difficulty associated with sentence difficulty. So, extraction of linguistic features from texts is a core task in the creation of automatic readability classifiers. This paper presents a

system that automatically extracts text features for European Portuguese and creates with them an automatic readability classifier. To accomplish this, it uses existing Natural Language Processing (NLP) tools, a parser and an hyphenator, and a corpora previously annotated by readability level. Currently, the system extracts 52 grouped in 7 groups: parts-of-speech (POS), syllables, words, chunks and phrases, averages and frequencies, and some extra features.

Two experiments were carried out to evaluate the classification task: one based on a five-level scale, taken from the Framework for Teaching Portuguese Abroad (in Portuguese, Quadro de Referência para o Ensino de Português no Estrangeiro, QuaREPE), published by the Ministry of Education and Science, and a second experiment based in a simplified three-level scale.

First, some related work is presented (Section 2), then the Natural Language Processing (NLP) tools here used (Section 3), followed by the features extracted from the text (Section 4), the automatic readability classifier developed (Section 5), the evaluation (Section 6) and, finally, the future work (Section 7).

2 Related work

There are several works on the topic of measurements and feature extraction for predicting the readability of documents. For English, early approaches consisted only in measuring simple features like the average sentence length, average number of syllables per word, *etc.* These methods include the *Flesch Reading Ease* [9], the *Fog Index* [13, 14], the *Fry Graph* [10] and the *SMOG* ("Simple Measure of Gobbleygook") [20] metrics. In general, these methods do not take into account the content of documents. It was not until later that content was taken into consideration in readability metrics, when some methods use a predetermined list of words to predict the reading difficulty, such as the *Lexile* [26] measure. More recently, language models have been used instead for this task, as in *Collins-Thompson & Callan* [27], where unigram language models were trained to predict the reading difficulty of English documents. Other methods, like *Schwarm and Ostendorf* [25], used syntactic features in addition to the language models, while some approaches, such as *Pitler and Nenkova* [22], relied on a variety of linguistic features, namely lexical, syntactic and discourse relations, in order to improve the classification.

Regarding the systems developed for Portuguese that are able to assess the readability of texts based on features extraction, one can refer REAP.PT[1] [19, 21] ("REAder-specific Practice for Portuguese"), a tutoring system for vocabulary learning (European Portuguese), which has been developed from the REAP system [5, 6] (English). Its readability measurement task is based on lexical features, such as statistics of word unigrams. Coh-Metrix-Port[2] [24] is yet another system that calculates parameters for measuring the cohesion, the coherence and the difficulty of a text. It was developed for the Brazilian Portuguese and has been adapted from the Coh-Metrix [12] (English).

[1] http://call.l2f.inesc-id.pt/reap.public (accessed in Feb 2014).
[2] http://www.nilc.icmc.usp.br:3000 (accessed in Feb 2014).

3 Natural language processing tools

To aid the extraction of features from European Portuguese texts, the system uses the natural language processing chain STRING[3] (Statistical and Rule-Based Natural Language Processing chain) [17] to extract statistical information about the texts. The number of syllables is extracted using the hyphenator YAH (Yet Another Hyphenator) [8].

The STRING [17] is an hybrid statistical and rule-based natural language processing (NLP) chain for Portuguese, that has been developed by L2F-Spoken Language Laboratory, at INESC-ID Lisboa. STRING has modular structure and performs all the basic NLP tasks, namely tokenization and text segmentation, part-of-speech tagging, morphosyntactic disambiguation, shallow parsing (chunking) and deep parsing (dependency extraction). STRING performs Named Entity Recognition, Information Retrieval, Anaphora Resolution and other NLP tasks and is composed of several modules, including a tokenizer, a morphological analyzer *LexMan* [28], morphosyntactic disambiguator called *RuDriCo* [7], a statistical POS tagger *MARv* [23], and a parser *XIP*[4] [15] (Xerox Incremental Parser).

The YAH Hyphenator [8] is a tool that has been developed by L2F-Spoken Language Laboratory, at INESC-ID Lisboa, designed by Ricardo Ribeiro and later improved by Figueirinha (2013) and is a rule-based system that applies various word processing division rules.

4 Features

The feature set extracted by the system consists in: (i) part-of-speech (POS), chunks, words and sentences features; (ii) verb features and different metrics involving averages and frequencies; (iii) several metrics involving syllable and (iv) extra features.

The features of group (i) are extracted from the chunking tree generated by STRING; features from groups (ii) and (iv) are also extracted from the chunking tree, but complemented by the dependencies' information generated by the processing chain; the metrics related to syllables (iii) are extracted using YAH.

Part-of-speech The system extracts the following POS categories: adjectives, adverbs, articles, conjunctions, interjections, nouns (common or proper), numerals, past participles, prepositions, pronouns (several subcategories), punctuation and special symbols.

The special symbols are, for example, "$", "%", "#", *etc.*

With this information extracted, the system calculates the POS relative percentages used in the readability assessment task. For example, conceptual information is often introduced through nouns and named entities, *e.g.* people's

[3]https://string.l2f.inesc-id.pt, (accessed in Feb 2014).

[4]Reference Guide: `https://open.xerox.com/Repo/service/XIPParser/public/XIPReferenceGuide.pdf` (last access: Feb. 2014).

names, locations, organizations, *etc.* These are important in text comprehension, yet the more entities and types of entities a text has, the harder it is to keep track of them and of the relations between all of them.

Chunks The system extracts the following chunks: nominal (NP), adjectival (AP), prepositional (PP) and adverbial (ADVP) phrases; temporal (VTEMP), aspectual (VASP) and modal (VMOD) auxiliary verb phrases; copulative (VCOP) verb phrases; past participle (PASTPART), gerundive (VGER) and infinitive (VINF) verb phrases; finite verb phrases (VF); and sub-clause phrases (SC and REL).

The information extracted allows the grouping of the chunks' relative percentages used in the readability assessment task. For example, the SC/REL type of chunks may be related to sentence hypotaxis complexity. According to the literature, the use of parataxis is preferable to a hypotactic structure, since a coordinated construction is in principle more easy–to–read and comprehensible than a subordinate one [3].

Sentences and words The system extracts the following features related to words and sentences: number of words, number of different words, number of sentences and word frequency.

The first three features are being used to calculate the averages and frequencies group. The length of a text is related with its readability, *e.g.* typically, long sentences have much more detail or content, which can make it more difficult for the readers to understand them.

The word frequency is related to the vocabulary used and, according to *Collins-Thompson & Callan's* approach, it is important to the readability assessment task. The linguistic motivation for using this parameter is that texts with more familiar vocabulary are easier to understand by the reader. The word frequency has been captured according to a language model based on unigrams, where the log-likelihood of a text is defined by the following expression:

$$\sum_{w} C(w) \times log(P(w|M)) \tag{1}$$

where P(w|M) is the probability of word w according to a background corpus M, and C(w) is the number of times w appears in the text.

This model will be biased in favor of shorter articles. Since each word has probability less than 1, the log probability of each word is less than 0, and hence including additional words decreases the log-likelihood. To overcome this issue, the system calculates this probability in n groups of 50 words each and then calculates an average of the n results.

The calculations are performed based on a set of distinct European Portuguese corpus provided by the AC/DC project available at Linguateca[5], using Laplace smoothing over the word frequencies.

[5]Distributed Resource Center for Computational Processing of Portuguese, http://www.linguateca.pt (accessed in April 2013)

Verbs In the verbs group, the system extracts the following features: number of different verbs, number of auxiliary verbs, number of main verbs and size of the verbal chains.

The system considers different inflections of the same verb as independent counts, since information clustered in this way is more interesting from a readability assessment point of view. The number of different verbs are extracted from the dependencies *VDOMAIN* and *VLINK*. The *VDOMAIN* is a binary dependency that links the first and last verb of a verb chain. In the case of a stand-alone verb, this is repeated in each argument of the *VDOMAIN* dependency. The *VLINK* is a binary dependency that links two consecutive verbs of a verb chain. For example:

Example 1. O Pedro leu o jornal 'Pedro has read the newspaper'

Example 2. O Pedro tinha começado a ler o jornal 'Peter had begun reading the newspaper'

in sentence 1, the stand-alone verb yields the dependency *VDOMAIN(leu,leu)* and there is no *VLINK* dependency. In sentence 2, the parser produces the dependencies *VDOMAIN(tinha,ler)*, on one hand, and *VLINK(tinha,começado)* and *VLINK(começado,ler)*, on the other hand. The use of different verbs and tenses increases the difficulty of the text.

Auxiliary verb constructions [2] are captured in different types of verb phrases' chunks, depending on their most proeminent grammatical value (temporal, aspectual or modal). For example, in sentence 2 *tinha* 'had' and *começado* 'begun' are parsed as auxiliaries, a temporal (VTEMP) and an aspectual (VASP) auxiliary verb phrases, respectively, of the infinitive (VINF) main verb chunk whose head is *ler* 'read'.

Auxiliary verbs may add complexity to the text, since they add functional or grammatical meaning to the clause in which they appear, for example, to express tense, aspect, modality, voice, *etc.*

The size of the verbal chains are extracted using the *VHEAD* dependency, which links each verb to the main verb of the verbal chain it belongs to. In the example 2, there are three such dependencies: *VHEAD(ler,ler)*, *VHEAD(começado, ler)* and *VHEAD(tinha,ler)*. Therefore, in this example, there is only one verbal chain with the size '3', since the head is the verb *ler* and the other two verbs are connected to it. The average size of the verbal chains can be useful for the readability measurement, since texts with shorter verbal chains can be easier to understand.

Averages and frequencies The system extracts the following averages and frequencies: average number of verb phrases per sentence, average length of sentences, average length of syllables per word, frequency of nouns and frequency of verbs. The frequency of nouns is the ratio of the number of nouns per number of words, and a similar ratio is calculated for the verbs.

The first two average measurements derive from *Pitler and Nenkova* approach [22], where it is pointed that the more verbs a sentence contains and

the longer a sentence is, the more complicated it becomes to understand it. The average length of syllables per word is important for readability measurement according *Flesch Reading Ease* and others metrics previously described 2. Words with higher length of syllables increases the difficulty of the text. On the frequency of nouns and verbs, *Coh-Metrix-Port* system [24] showed that higher frequencies translate to harder readability, distinguishing complex texts (adults) from simple texts (children).

Syllables The system extracts the total number of syllables and the ratio words/syllables.

Extras The system extracts also the total number of dependencies, total number of tree nodes, number of pronouns per noun phrases (NP), number of NP with a definite or demonstrative determiner, number of NP with a indefinite determiner, number of subordinate clauses (SC chunks), number of coordination relations, size of coordination relations' chains and readability measure (Flesch Reading Ease BR).

The first five features are extracted from the POS and chunks groups. The number of pronouns per noun phrases derive from *Coh-Metrix-Port* system [24]. The greater the number of pronouns per noun phrases, the more difficult it becomes to identify who or what the pronoun refers to. The idea of extracting the number of NP with a definite or demonstrative determiner is related with the fact that, usually, the presence of those determiners imply a reference to a previous words, as opposed to indefinite determiners. The text with lowest definite/indefinite NP ratio should be more cohesive, since the presence of NP with a definite or demonstrative determiner involve anaphora processing.

The number of subordinate clauses is extracted from several dependencies with the feature *SENTENTIAL* or *RELAT* (for relative sub-clauses). For example, the *MOD* is a binary dependency that links a modifier with the element it modifies. The *SENTENTIAL* feature indicates that the modifier is a sub-clause and, in this case, it links the main verb of the main clause to the main verb of the sub-clause. For example, in sentence 3:

Example 3. O Pedro desconfiava do facto de a Ana ter ido a Lisboa 'Pedro suspected from the fact that Ana had gone to Lisbon'

the following dependency is extracted *MOD_SENTENTIAL(desconfiava,ido)*. Other dependencies besides *MOD* can get the *SENTENTIAL* feature, most prominently *SUBJ* (subject) and *CDIR* (direct object). *RELAT* also involves a subordinated clause, but this is a relative clause, and it modifies the head of the noun phrase it belongs to[6]. For example, in sentence 4:

Example 4. A moça que era cozinheira no palácio fazia bolos 'The girl who was a cook in the palace made cakes'

[6]A similar relation is also established in the case of appositive/explicative relative sub-clauses.

the following dependency is extracted *MOD_RELAT(moça,era)*. The number of subordinate clauses are related with the hypotaxis complexity previously mentioned.

The number of coordination relations is extracted from the *COORD* dependency. The *COORD* dependency establishes a coordination relation between elements of a coordination chain. For example, in sentence 5:

Example 5. O Pedro comprou uma pera e duas laranjas, enquanto a Maria comprou uma sopa, uma sandes e um sumo 'Pedro bought two oranges and a pear, while Maria bought a soup, a sandwich and a soda'

two coordination chains are present: on one hand *COORD(e,pera)* and *COORD(e,laranjas)*, and, on the other hand, *COORD(e,sopa)*, *COORD(e,sandes)* and *COORD(e,sumo)*.

The size of coordination relations chains are obtained from *CLINK* dependency. The *CLINK* dependency defines the link between two consecutive words in a coordinated chain. In the example 5, the *CLINK* dependencies extracted are *CLINK(pera, laranjas)*, for the first chain, and *CLINK(sopa, sandes)* and *CLINK(sandes, sumo)*, for the second chain. With this information, the system pairs the results from *CLINK* and retrieves the size of all the coordination chains' relations: two (*pera* and *laranjas*) and three (*sopa*, *sandes* and *sumo*). The number of coordination relations and the length of their chains are related with the parataxis complexity (complementary of hypotaxis).

5 Readability Classifier

According to the Framework for Teaching Portuguese Abroad (in Portuguese, Quadro de Referência para o Ensino de Português no Estrangeiro, QuaREPE), published by the Ministry of Education and Science, it is considered that the degree of proficiency in a foreign language can be determined on a scale of five levels [18]:

- A1: initiation;
- A2: elementary;
- B1: intermediate;
- B2: upper intermediate;
- C1: advanced.

The classification task has two experiments, one based on this five-level scale and a second experiment based in a simplified three-level scale, *i.e.*, the classifier is trained to predict if the text belongs to level A, B or C. We consider the second experiment useful because distinguishing between the levels A1 and A2; B1 and B2 may be very difficult.

For classification proposes, we tested several machine learning algorithms available in WEKA machine learning tool[7] [4] (Table 1).

[7] http://www.cs.waikato.ac.nz/ml/weka (accessed in April 2014).

Table 1. Weka's learning algorithms tested

Learning method	Algorithms
Bayes	Naive Bayes
Linear	Support Vector Machines (SVM) Logistic regression
Lazy	K-nearest neighbors learner (IBk) K* (KStar)
Boosting	AdaBoost LogitBoost
Rules	Holte's OneR
Decision tree	C4.5 (J48) C4.5 grafted (J48graft) Decision stumps Random Forest

5.1 Corpus

The corpus used to train the classifier consists of a set of 212 texts, previously classified in regards to their intelligibility and provided by the Instituto Camões[8]. This corpus was created from tests, exams and materials used for the teaching of European Portuguese. The manual classification of the intelligibility of texts takes into account reading/comprehension skills stipulated by the *Framework for Teaching Portuguese Abroad* published by the Ministry of Education and Science for each level[9]. Table 2 shows the corpus distribution for each readability level.

Table 2. Corpus distribution.

	A1	A2	B1	B2	C1
# Texts	23	33	128	12	16
Percentage	11%	16%	60%	6%	7%

[8] http://www.instituto-camoes.pt (accessed in June 2014)

[9] `http://www.dgidc.min-edu.pt/outrosprojetos/data/outrosprojectos/Portugues/Documentos/manual_quarepe_orientador_versao_final_janeiro_2012.pdf` (accessed in April 2014).

6 Evaluation

6.1 Feature extraction system

The text used for evaluation of the feature extraction system was extracted from a journalistic text that is part of the European LE-PAROLE Portuguese corpus [1]. This text has 490 words and 14 sentences. Table 3 presents an analysis of the system results, decomposed by features groups.

Table 3. Results for the evaluation of features group extracted.

Feature group	Found	Correct	Reference	Precision	Recall	F-measure
POS	576	553	556	96.01%	99.46%	97.70%
Chunks	269	264	271	98.14%	97.42%	97.78%
Verbs	153	153	159	100%	96.23%	98.08%
Sentences and words	796	796	796	100%	100%	100%
Averages and frequencies	39.57	39.52	39.52	99.87%	100%	99.94%
Syllables	1245	1236	1255	99.28%	98.49%	98.88%
Extras	45.65	45.64	45.64	99.99%	100%	99.99%
Total	3134.22	3087.16	3122.16	98.81%	98.88%	98.85%

Note 1. "Found", "Correct" and "Reference" means, respectively, the number of features identified by the system, features correctly identified by the system and features manually annotated in the corpus.

Table 3 shows the system efficiency in correctly identifying most of the features (f-measure from 97.70% to 100%). The POS feature group has the lowest precision, since the system found 158 nouns, where the annotated corpus has only 144. The lowest recall was in the verb group where 56 out of 63 verbs were identified. This incomplete identification of the verbs had impact on the chunk group's recall. The system attained an f-measure of 98.85%.

In a complementary approach, the system performance has been measured using 12 journalistic texts, with different sizes, from the same corpus used on the feature evaluation. Table 4 presents the results of the performance evaluation.

These results show that although the system's performance depends on the text's size, a 1200% increase on the number of words took only around 40% more CPU time. STRING is the component of the system that requires more time to process, it as expected, due to all the tasks it performs.

6.2 Feature contribution

To assess the contribution of the factors extracted in readability classification, we used the WEKA toolkit with the feature selection algorithm *InfoGainAttributeEval*[10]. This evaluation was conducted in the two different scenarios previously

[10] `http://weka.sourceforge.net/doc.stable/weka/attributeSelection/InfoGainAttributeEval.html` (accessed in June 2014).

Table 4. Results for the system's performance with different text sizes.

Text size (words)	Feature extraction (s)	STRING (s)	YAH (s)
88	1.421	7.393	0.004
144	1.598	7.573	0.005
199	1.675	7.690	0.007
290	1.977	7.833	0.009
332	1.985	7.943	0.009
397	2.101	8.017	0.011
499	2.220	8.353	0.015
534	2.297	8.453	0.013
598	2.421	8.603	0.013
649	2.393	8.677	0.016
900	2.516	9.300	0.020
1065	2.659	9.857	0.024

Note 2. The performance has been calculated (in seconds) in user CPU time. The results presented are averages of 3 executions per text.

mentioned (Section 5). Figures 1 and 2 show the results for the features with higher contribution on the classification task.

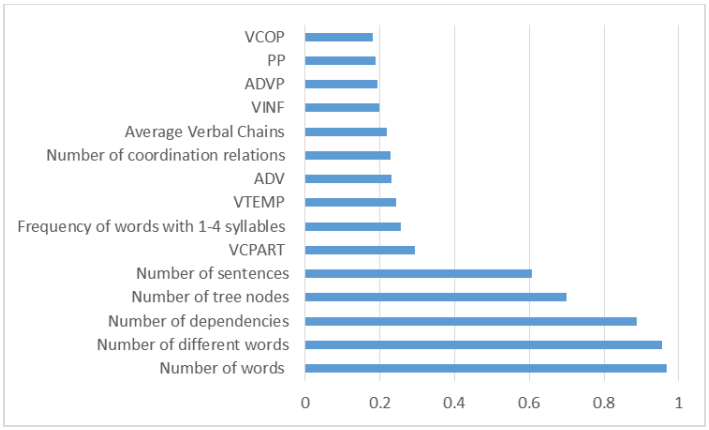

Fig. 1. Feature contribution for the five levels scale classification.

We conclude that, in the two scenarios, the features that most influence the classification were the number of words, number of different words, number of dependencies, number of tree nodes and number of sentences. In the next features, we see that the frequency of words with lowest number of syllables, frequency of adverbs and average verbal chains were important features for the classification in both scenarios. From this point forward, we see that the features

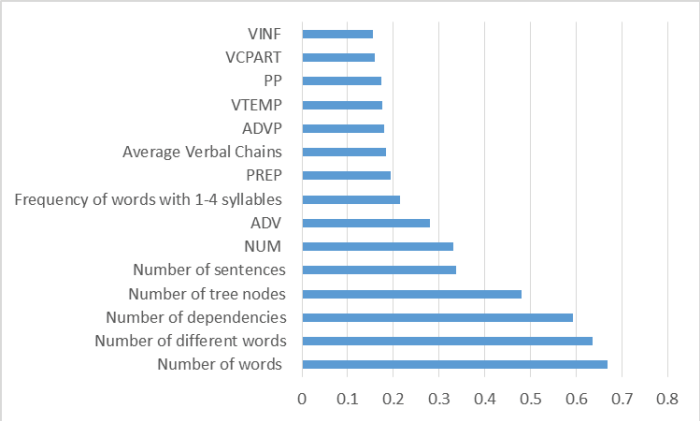

Fig. 2. Feature contribution for the three levels scale classification.

contributions significantly differ between the two scenarios presented. However the rest of the features were related with some chunks and POS frequencies. The results here presented are only focused on fifteen features that stood out in the evaluation. However, the classifier uses fifty-two features.

6.3 Readability classifier

In both scenarios, several machine learning algorithms available in WEKA machine learning toolkit were tested and the algorithm with best results was the LogitBoost (Table 7 and 10). The evaluation was performed using 10-fold cross-validation. The metrics chosen for analyzing the quality of the classifier were accuracy (percentage of correctly classified instances), root mean square error (RMSE), ROC Area and Kappa statistics. Additionally a confusion matrix and algorithm performance comparison is presented for each scenario.

Five level classification In this scenario, we also present the *adjacent accuracy within 1 grade level*. This is the percentage of predictions that are equal to or show one level of difference to the manually assigned label. Measuring strict accuracy is considered too demanding because manually assigned labels are not always consistent.

Table 5. Evaluation of the readability classifier (five levels).

	Accuracy	**RMSE**	**ROC Area**	**Kappa**	**Adjacent Acc.**
Cross-validation	79.25%	0.246	0.948	0.638	0.92

Table 6. Confusion Matrix (five levels).

		Predicted class				
		A1	**A2**	**B1**	**B2**	**C1**
Actual class	**A1**	12	5	6	0	0
	A2	6	20	7	0	0
	B1	2	2	122	2	0
	B2	0	1	1	5	5
	C1	0	1	0	6	9

In this scenario, the classifier correctly classified 79.25% instances, *e.g.*, 168 texts (Table 5). It is interesting to notice that for most texts, the assigned level is either correct or mostly within one-level difference (Table 6). As expected, the adjacent accuracy is 0.92 and the RMSE result is low because the values expected and the values observed are close. The Kappa is a chance-corrected measure of agreement between the classifications and the expected values, where 1.0 represents perfect agreement. It will be useful to compare this experiment with the scenario below.

Table 7. Algorithms comparison results (five levels classifier)

Algorithms	**Accuracy**	**RMSE**
Naive Bayes	74.06%	0.307
Support Vector Machines	77.83%	0.333
Logistic regression	67.45%	0.356
K-nearest neighbors learner	72.17%	0.330
K*	70.76%	0.327
AdaBoost	65.57%	0.333
LogitBoost	**79.25%**	**0.246**
Holte's OneR	70.28%	0.345
C4.5	74.06%	0.308
C4.5 grafted	76.42%	0.293
Decision stumps	65.57%	0.286
Random Forest	78.30%	0.251

Three level classification In this scenario, the *adjacent accuracy within 1 grade level* is not calculated, because there are only three levels and the level B will have always the maximum value. The three level classification obtained RMSE and ROC area values similar to the previously mentioned classifier and achieved 86.32% of accuracy (Table 8). However, it has a highest Kappa value than the five level classification, which indicates that this classifier has a better agreement.

Table 8. Evaluation of the readability classifier (three levels).

	Accuracy	RMSE	ROC Area	Kappa
Cross-validation	86.32%	0.265	0.933	0.7221

Table 9. Confusion Matrix (three levels).

		Predicted class		
		A	**B**	**C**
Actual class	**A**	49	7	0
	B	8	125	7
	C	1	6	9

Table 10. Algorithms comparison results (three levels classifier)

Algorithms	**Accuracy**	**RMSE**
Naive Bayes	80.19%	0.353
Support Vector Machines	82.55%	0.343
Logistic regression	75.94%	0.392
K-nearest neighbors learner	76.89%	0.390
K*	78.30%	0.365
AdaBoost	70.76%	0.333
LogitBoost	**86.32%**	**0.265**
Holte's OneR	75.00%	0.408
C4.5	80.19%	0.359
C4.5 grafted	80.19%	0.336
Decision stumps	72.64%	0.336
Random Forest	84.43%	0.283

7 Future work

Additional features can be extracted in order to improve the system presented, namely the number of omitted subjects, level of hypotaxis/sentence, among others. For example, the number of omitted subjects can be extracted from all dependencies that have the *ELIPS* or *ANAPH0* feature. The *ELIPS* feature indicates that a subject noun phrase of the main verb of the sentence has been zeroed and that it has been reconstructed based on the person-number agreement. For example, in sentence 6:

Example 6. Vamos hoje ao cinema '[We] are going to the movies today'

an elliptic subject has been reconstructed as *SUBJ_ELIPS*(Vamos,Nós), based of the 1st.person-plural inflection of the verb. The *ANAPH0* feature indicates that a subject of a sub-clause has been zeroed (zero anaphora) and it has been reconstructed from a previously mentioned antecedent. For example, in sentence 7:

Example 7. A Joana comprou um livro e leu-o 'Joana bought a book and has read it'

the following anaphoric subject dependency is produced: *SUBJ_ANAPH0*(leu, Joana). The subject omission occurs mostly when the subject as already been presented in the same or in the previous sentence though it also may depend on other factors, such as the main verb of main clause, the subordinate conjunction, among others.

Subject ellipsis, and other types of zero anaphora, can complicate the interpretation of a text, specially for readers that are starting to learn a new language, so it should be taken into consideration for text readability assessment.

8 Conclusions

This paper presented a classifier for European Portuguese texts based on a variety of linguistic features. It seeks to assist the selection of adequate reading materials for teaching European Portuguese as a second language for different language proficiency levels.

Associating readability scores to texts is also important in other areas, such as in the production of medical information, tools and software manuals, safety instructions, *etc.*, whose correct interpretation is essential to avoid different types of risk and to make such texts accessible reading to the majority of the population.

The system here presented, focused on 52 grouped in 7 groups with an f-measure of 98.85%. These features are helpful to evaluate the readability of the texts as showed by the results presented, highlighting the number of words, number of different words, number of dependencies, number of tree nodes and number of sentences, frequency of words with lowest number of syllables, frequency of adverbs, average length verbal chains and some chunks and POS frequencies.

In both scenarios, with five readability levels (A1 to C1) or with three levels (A, B or C), the classifier here developed achieved good results with an accuracy of 79.25% and 86.32%, respectively, and most of the errors are within one-level distance from the expected results.

In the future, the system here presented will be made available to the general public through a web form and it can easily be extended by adding new features or metrics of interest to the task at hand.

References

1. Bacelar do Nascimento, M.F., Marrafa, P., Pereira, L.A.S., Ribeiro, R., Veloso, R., Wittmann, L.: LE-PAROLE-Do corpus à modelização da informação lexical num sistema-multifunção. Actas do XIII Encontro da Associação Portuguesa de Linguística (1998) 115–134
2. Baptista, J., Mamede, N., Gomes, F.: Auxiliary verbs and verbal chains in European Portuguese. In: Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language. PROPOR'10, Berlin, Heidelberg, Springer-Verlag (2010) 110–119
3. Beaman, K.: Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. In Tannen, D., ed.: Coherence in Spoken and Written Discourse. Volume 12. Ablex, Norwood, NJ (1984) 45–80
4. Bouckaert, R.R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., Scuse, D.: WEKA manual for version 3-7-8. (2013)
5. Brown, J., Eskenazi, M.: Retrieval of authentic documents for reader-specific lexical practice. In: Proceedings of InSTIL/ICALL Symposium 2004. Volume 17., Venice, Italy (2004)
6. Collins-Thompson, K., Callan, J.: Information retrieval for language tutoring: An overview of the REAP project. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '04, New York, NY, USA, ACM (2004) 544–545
7. Diniz, C., Mamede, N.J., Pereira, J.C.S.D.: RuDriCo2 - a faster disambiguator and segmentation modifier. In: II Simpósio de Informática (INForum), Universidade do Minho (September 2010) 573–584
8. Figueirinha, P.: Syntactic REAP.PT. Exercises on word formation. Master's thesis, Instituto Superior Técnico - Universidade de Lisboa (October 2013)
9. Flesch, R.: Marks of Readable Style: A Study in Adult Education. Number no. 897 in Contributions to education. Columbia University, Teachers College, Bureau of Publications, New York, United States (1943)
10. Fry, E.: A readability formula that saves time. Journal of Reading **11**(7) (1968) 513–578
11. Fulcher, G.: Text difficulty and accessibility: Reading formulae and expert judgement. System **25**(4) (1997) 497 – 513
12. Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z.: Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments, & Computers **36**(2) (2004) 193–202
13. Gunning, R.: The Technique of Clear Writing. McGraw-Hill, New York, USA (1952)
14. Gunning, R.: The Fog Index after twenty years. Journal of Business Communication **6**(2) (1969) 3–13

15. Hagège, C., Baptista, J., Mamede, N.J.: Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O segundo HAREM. (January 2008)
16. Klare, G.: The measurement of readability. Iowa State University Press, Ames, USA (1963)
17. Mamede, N.J., Baptista, J., Diniz, C., Cabarrão, V.: STRING: An hybrid statistical and rule-based natural language processing chain for Portuguese. In: International Conference on Computational Processing of Portuguese (Propor 2012). Volume Demo Session., Coimbra, Portugal (April 2012)
18. Maria José Grosso, António Soares, F.d.S.J.P.: QuaREPE - Quadro de Referência para o Ensino de Português no Estrangeiro. Documento Orientador. Ministério da Educação/Direção Geral de Inovação e Desenvolvimento Curricular, Lisboa. (2011)
19. Marujo, L., Lopes, J., Mamede, N.J., Trancoso, I., Pino, J., Eskenazi, M., Baptista, J., Viana, C.: Porting REAP to European Portuguese. In: ISCA International Workshop on Speech and Language Technology in Education (SLaTE 2009), Wroxall Abbey Estate, Warwickshire, England (September 2009)
20. McLaughlin, G.H.: SMOG grading: A new readability formula. Journal of Reading **12**(8) (1969) 639–646
21. Pellegrini, T., Ling, W., Silva, A., Correia, R., Trancoso, I., Baptista, J., Mamede, N.J.: Overview of computer-assisted language learning for European Portuguese at L2F. In Helfert, M., Martins, M.J., Cordeiro, J., eds.: CSEDU (2), Porto, Portugal, SciTePress (2012) 538–543
22. Pitler, E., Nenkova, A.: Revisiting readability: a unified framework for predicting text quality. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP '08, Stroudsburg, PA, USA, Association for Computational Linguistics (2008) 186–195
23. Ribeiro, R.: Anotação morfossintátictica desambiguada do Português. Master's thesis, Instituto Superior Técnico (March 2003)
24. Scarton, C.E., Aluísio, S.M.: Análise da inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. Linguamática **2**(1) (2010) 45–61
25. Schwarm, S.E., Ostendorf, M.: Reading level assessment using support vector machines and statistical language models. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. ACL '05, Stroudsburg, PA, USA, Association for Computational Linguistics (2005) 523–530
26. Stenner, A.J.: Measuring reading comprehension with the Lexile framework. In: Fourth North American Conference on Adolescent/Adult, London, UK, Academic Press Ltd (1996)
27. Thompson, K.C., Callan, J.P.: A language modeling approach to predicting reading difficulty. In: HLT-NAACL, Boston, United States (2004) 193–200
28. Vicente, A.M.F.: LexMan: um segmentador e analisador morfológico com transdutores. Master's thesis, Instituto Superior Técnico, Universidade de Lisboa (June 2013)