Assessing Automatic Text Classification for Interactive Language Learning

António Branco, João Rodrigues, Francisco Costa, João Silva
Department of Informatics, Faculty of Sciences
University of Lisbon
Portugal

Rui Vaz
Divisão de Programação, Formação e Certificação
Camões IP
Portugal

Abstract— **In this paper we discuss the design options for a language processing tool that supports humans in their task of classifying text excerpts according to CEFR levels of language proficiency. We describe the tool that we developed on the basis of these design options and provide an assessment of its functioning. This tool is suitable to be used by students taking courses of Portuguese as a second language, as well as by expert instructors selecting items of exam that are aimed at assessing and certifying the language level of these students. It is an instrument that aims at supporting humans in their language level classification judgments by providing conditions to more consistent and objectively sustained judgments across different occasions and input texts and across different human classifiers. Its design principles and underlying language technology can be applied to develop similar tools for any other language.**

Keywords — language learning assessment; readability assessment; CEFR; Portuguese language.

I. INTRODUCTION

The potential of eLearning continues to be deployed, with the recent outburst of interest in massive open online courses (MOOCs) being one of the more visible advancements in the area. This interest spans also to natural language, either with with the expanding of the scope of MOOCs to language learning (e.g. Instreamia, 2014) or even with the learning management systems resorting to language technology tools (e.g. Monachesi *et al.*, 2006; Avelãs *et al.*, 2008). The present paper encompasses both these two aspects as we discuss how, in the context of language learning courses, natural language processing tools can be fruitfully applied to support these courses.

With massive online courses, a range of new challenges emerges. In this paper we concentrate on the demands concerning the certification of the skills and knowledge acquired by students. Our specific focus is on the elaboration of quizzes and exam items for students who are learning a second language.

In more concrete terms, we are considering the certification provided by Camões IP with respect to language level attained by students learning Portuguese as a second language (Camões, 2014). A major stressing demand in the organization of a certification process like this is the elaboration of quizzes and exam items. Given the very large number of students and certification events being handled, the pool of such materials ready to be used needs to be permanently renewed, given that for obvious reasons of the integrity of the evaluation process, each exam item can be used only once to test students' knowledge.

In this respect, an important need consists in finding appropriate text excerpts that can be used and quoted in the exams and in quizzes, and upon which questions and exercises to be solved by the students can be drawn. Given that we will be considering five of the CEFR[1] language levels for certification (levels A1, A2, B1, B2 and C), a key difficulty relies on finding excerpts of appropriate levels of language. This difficulty of course adds to the demand that these excerpts need to be "naturally occurring" ones, that is they cannot be constructed by the instructor authoring the exam. And it adds also to the demand that their renewal rate is as high as the renewal rate of the items in the pool, with each excerpt being used only once in one of the exams of the pool.

In short, when running massive language learning courses, one of the important back office demands is thus the task of continuously finding new text excerpts, classified according to language levels, to support a pool of quizzes and exams in permanent renewal.

The goal of this paper is to discuss how natural language technology can be called upon to support this task, and thus how it can help make it possible in practice to run massive language learning courses for Portuguese, and for that matter in any language.

In particular, given the current state of the art in the computational processing of Portuguese, we describe a tool, available as an online service, which we developed to be used by the authors of the quizzes and exams to support their productivity in determining whether excerpts are appropriate for the language level at stake.

As this tool was commissioned by Camões IP, and was made available as an online service (http://lxcefr.di.fc.ul.pt), it became nevertheless useful both for instructors and students. It helps to improve the classification of candidate excerpts by the

[1] Common European Framework of Reference for Languages: Learning, Teaching, Assessment (Council of Europe, 2001; Conselho da Europa, 2001).

experts. And it helps also to enhance the level of interactivity of the language courses for students, as they can resort to this tool to check whether a new text they come across may be appropriate for their current language level.

The present paper is organized as follows. In Section II, we put the identified needs in perspective, considering possible available solutions in terms of language technology. A brief overview of metrics and classifiers to automatically assess the readability of texts will be provided in Section III. The following Section IV discusses the design options with regards to these instruments in view of the purpose and practical constraints of the task at hand. Section V introduces the instruments selected and the tool developed, while Section VI discusses the technological aspects involved in its implementation. This is followed by the reporting on the evaluation of different dimensions of the tool, in Section VII. The last Sections VIII and IV offer discussion and conclusions.

II. LANGUAGE TECHNOLOGY OUTLOOK

The application to be implemented was aimed at being a tool that supports humans in their task of determining whether a given several sentence long text excerpt is appropriate to be used in a quiz for language learning or in an exam for language learning assessment. In particular, this tool should help to speed up humans in their making of the judgments, and support their reliability on these judgments, about whether given excerpts are appropriate for given levels of language, in a scale of five such levels of language. To facilitate its use and availability for end users, this tool should be available in the form of an online service accessible through web browsers.

At first blush, this kind of tool seems to come close to belonging to the area of Computer Assisted Language Learning (CALL). But at a closer inspection, it turns out in fact not to be a CALL application proper, or then it can be considered to be so only very remotely. It is aimed at supporting the making of quizzes and exams, at the instructors' end, and not at supporting the process of language learning per se, at the students' end.

It might seem also that this tool could belong to some kind of assistive technology, but in fact the scope of what is typically taken as assistive technologies encompasses solutions of quite a different nature, more related to some form of human disability.

When focusing on the functionality of this tool, it seems to address a problem similar to the well-known task of text categorization. But in fact, text categorization is about automatically associating texts to content domains, topics or subject matters, while here the excerpts of a same given language level can pertain to different domains.

This tool also seems to come close to the long researched problem of readability assessment, but again our problem seems to have something specific to itself, as it has a categorical nature, that is the excerpts are not expected to be located in a continuum, but in a specific scale of five separate levels of Portuguese language for which no objective or linguistic criteria have been made fully explicit and can be resorted to so far.

These considerations tend to indicate that while the technological solution for our problem may take inspiration from other similar problems and respective solutions, our tool emerges as having some specific nature or demands that may be different from the other well known applications based on natural language processing (Mitkov, 2003).

In this quest to envisage the best approach to deal with our working problem, it is important to bear in mind also that in the current state of the art, there are no data sets that would support resorting to machine learning techniques, as there is no data set of excerpts classified according to the relevant language levels with enough volume to support the use of this type of technology.

In spite of this, we believe that it is nevertheless possible to resort to other options in order to develop an instrument useful to support this human task: an instrument that helps to obtain better results, by offering conditions to more uniform and sustained classification judgments across different occasions and texts, and across different human classifiers.

As we will discuss in the remainder of this paper, this tool should take its primary inspiration from the work on indexes and classifiers for readability, and should resort to the support by language processing tools with already pretty good accuracy and performance, such as tokenizers, POS taggers, syntactic parsers, etc.

III. READABILITY INDEXES AND CLASSIFIERS

The assignment of a text excerpt to a language level certainly depends on many different factors impinging on that excerpt, including for instance expected audience, content, coherence, legibility, readability, etc. (Council of Europe, 2001). In our work here, we focus on readability, for which there are results from a long established tradition of research, with operational solutions widely used by official institutions all over the world (DuBay, 2004). We focus on readability also because natural language processing not only can help make automatic the calculation of indexes proposed in the past, but also because it is driving the research in this area into a new path of new promising results (cf. overview in Feng *et al.*, 2010, Dell'Orletta *et al.*, 2011).

There is a long tradition of research on instruments to support readability assessment, with a wide range of different readability indexes being defended (cf. overview in DuBay, 2004). In spite of the many indexes proposed along the years, the index proposed by Rudolf Flesch on the 1940's seems to continue to have the widest acceptance and usage. Also known as the Flesch Reading Ease index (Flesch, 1979), it is calculated with the following formula:

206.835 - 1.015(total words/total sentences) - 84.6 (total syllables/total words)

Higher scores indicate texts that are easier to read. The constants in the formula are just instrumental to bring the resulting scale between the scores 0 and 100 to be aligned with reading development categorized along the number of years in formal education as mother language, tuned to the USA context. Of course, what is worth noting here is that the core of the index relies on two ratios, one at the word and the other at

sentence level, that are language independent. These turn out to be quite simple ones and very straightforward to calculate, namely the average sentence length in terms of number of words, and the average word length in terms of number of syllables.

In practice, other metrics pretty much correlate with these two basic ones to some extent (DuBay, 2004), and in any language they can serve as a basis of comparison for excerpts of different levels of readability (no matter the actual range of figures produced).

In spite of the success of readability indexes, they have always been surrounded by controversy, mostly because these indicators are criticized on the assumption of goals they ultimately do not aim to fulfill. These indexes are meant to be indicators that support and need to be complemented with human judgment, which should take into account all the other relevant factors for the complexity of a text, as noted at the beginning of this section. They are auxiliary tools for human judgment, not accurate one-stop predictors.

Recently, research on readability assessment has gained popularity in the area of natural language processing. To a great extent, this is the result of the potential of using classifiers, trained on previously annotated material, to estimate the readability level of an input text, or just its level of simplification (for overview and experimental contrastive assessment, see Feng *et al.*, 2010).

In this respect, it is worth of special mention the work reported in (Aluísio *et al.*, 2010) for a couple of reasons: they worked with the Portuguese language; and this work is one of the few systematically exploring a wide experimental space, which involved 60 features and three types of classifiers.

While signaling the importance of the above results, it is nevertheless relevant to underline also the experimental differences with respect to the task we are concerned with here in this paper. The tool sought for in (Aluísio *et al.*, 2010) had to deal with only a ternary distinction (between "original text", "simplified text", and "strongly simplified text"). And very importantly, their training materials consisted of data sets that were manually constructed by means of simplification of an original natural occurring version, along grammatical dimensions for which there were features that the classifiers happened to be eventually trained upon.

In our case, the challenge is different and more demanding. Our tool has to handle larger number of levels of classification, that is five levels. Additionally, it has to cope with any naturally occurring input text where the text of a given level has no special constructed relation, in terms of content or of formal properties, with texts that may happen to be in the other levels.

But above all, the purpose of the application is different. The tool described in (Aluísio *et al.*, 2010) is due to feed a subsequent automatic simplification system. In our case, the outcome of the tool is due to help humans in their task of text selection for language learning and assessment.

IV. METRICS TO SUPPORT MANUAL CATEGORIZATION

As mentioned above, when seeking to explore the potential of language technology to design our tool, a first challenge is that we cannot count on pre-existing data with enough volume for developing sophisticated classifiers. Furthermore, given the nature of the difference between the five levels, and the typical size of a relevant excerpt, there is no obvious way to undertake the process of manually producing such a data set from existing data consisting of naturally occurring texts already classified (in a possible attempt to mimic (Aluísio *et al.*, 2010) on how they obtained their data). Given this, to construct our tool, we had to base it on shallow yet as reliable as possible metrics.

A second important challenge to take into account is the purpose of the tool and its utilization by human operators. An important constraint to take into account is thus that the indicators and values to be obtained by these tools have to be able to be interpreted by humans in view of their task at hand. Also, and very importantly, for these indicators to be useful, the instrumental tools extracting them have to be able to estimate them with good enough accuracy.

The quantitative metrics to take into account have to be manageable by humans in view of facilitating their classification judgment and making that process more rapid and agile. Hence, it is not viable to present as much as the whole set of 60 “features” handled by the system in (Aluísio *et al.*, 2010) or the 200 “measures” of the Coh-Metrix system (Graesser *et al.*, 2004), or even a subset of 30 selected measures of the Coh-Metrix system, as in (Scarton *et al.*, 2009). All these metrics may turn out to be important for subsequent automatic processes or human investigators possibly doing research on quantitative linguistics. But they would be counterproductive for someone seeking to perform the task of classifying excerpts according to language level as rapidly and as efficiently as possible. They would also be cumbersome as most of them correlate more or less with a few core other ones (DuBay, 2004; Feng *et al.*, 2010). It is thus worth of note that (Rodrigues *et al*., 2013), while pursuing a quite different goal, had to face the same need of focusing on a sensible subset of metrics.

The metrics need thus to be interpretable by humans. This leads to set aside a number of features used in the literature whose scores are hardly interpretable by themselves, e.g. the values for the “perplexity of trigrams” or the “probability of unigrams” in a language model, or “the number of high level syntactic constituents”, among several others.

Additionally, the metrics, and their calculation, need to rely on language processing tools and technologies whose state of the art offers good enough accuracy, so that in practice the error rates of these tools are not passed onto and do not undermine the reliability of the metrics for the human user. Just as an example: given the absence of a wide enough WorldNet for Portuguese, this should exclude taking into account a metric such as a “noun ambiguity ratio”, etc.

For our tool to be robust across different domains, the metrics to be used should also be independent of considerations concerning lexical content supported by indexes such as those relying on some predefined lexical lists or lexical selection.

Hence metrics based on lists of lexical items could not be considered.

Finally the metrics need to allow for normalized values that permit comparative assessment among different excerpts.

After considering these constraints, we focused on tools whose performance is reliably over 95% of accuracy, which given the relevant targeted metrics, encompasses tokenizers, sentence splitters, syllabifiers and POS taggers. Exceptionally, given the specific needs of counting different types of clauses, we took on board also statistical syntactic constituency parsers, the only type of tools we used whose performance in terms of the relevant metrics still scores below 90% according to their current state of the art.

V. QUANTITATIVE METRICS

Taking into account a careful pondering of the above considerations, we eventually selected four metrics, with one metric per each one of three grammatical levels, plus a fourth one, which combines factors from different dimensions:

Lexical dimension:

- Lexical category density in proportion of nouns

Word dimension:

- Average word length in number of syllables per word

Sentence dimension:

- Average sentence length in number of words per sentence

Combined:

- Flesch index

For each dimension above, we picked the metric that in the literature is typically pointed out as the one having the best predictive power. For instance, Feng *et al*. (2010) indicate that "in general, [lexical category density metrics] appear to be more correlated to text complexity than syntactic features, shallow features and most discourse features" (p.283), underlining that "among the five word classes investigated, noun-based features generate the highest classification accuracy" (p.282). They also indicate that "experimental results […] show that average sentence length has dominating predictive power over all other shallow features".

As depicted in Table 1, the data set we had available to work with is very small. This data set is composed of excerpts that occurred in the exams that were used in 2013 by the certification services of Camões IC to perform language level assessment across the world for learners of Portuguese as second language. These excerpts were selected by human experts with the purpose of being integrated in these exams, and were thus carefully classified as belonging to one of the five language levels, the level that corresponded to the level of the exam where they were eventually integrated.

TABLE I. STATISTICS OF THE DATA SET AVAILABLE

Level	Excerpts	Tokens	Av. tokens /excerpts	Sentences	Av. sentences /excerpts
A1	11	1122	102.00	74	6.73
A2	11	1730	157.27	117	10.64
B1	68	4142	60.91	159	2.34
B2	8	2472	309.00	149	18.63
C1	12	3207	267.25	172	14.33
Total	110	12673	115.21	671	6.10

As mentioned above, the small volume and high unbalance (e.g. there are 68 B1 texts against only 8 B2 ones) of this data set does not permit to expect a reliable use of machine learning techniques and text classifiers. Furthermore, it was important to seek to elucidate whether these data were even sufficient to support the shallower metrics we should use. For each of the three dimensions above (lexical, word, sentence), we wanted to check the impact of the volume of the data available to support relevant correlations, and thus to eventually support the reference scale needed to be set up.

The results are displayed in the charts in Fig. 1.

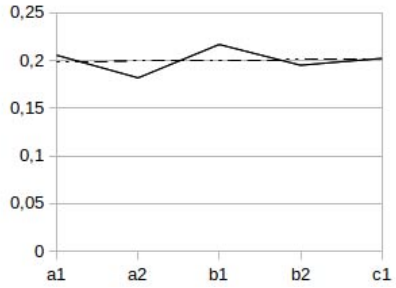
As can be seen from the linear regression over each chart, the data set available, though very small, already indicates an overall tendency of supporting a correlation between larger scores for the metrics and more advanced language levels, in all metrics except the Flesch index, where as expected this direction of the correlation is inverted. The metric that is yet little responsive, given the volume of data available we hope, is the lexical metric based on the proportion of nouns (but see also further discussion on this below, in Subsection VII.A).

Reference scales were drawn on the basis of the average scores of these metrics for each language level. For ergonomic reasons, these reference scores are displayed in a radar chart, as exemplified in Fig. 2, after being projected into linear scales.

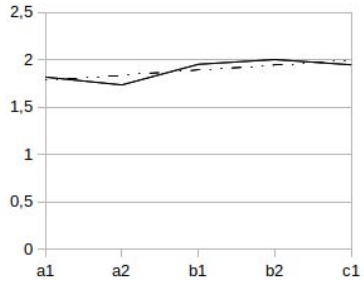
As exemplified in that figure, the scores obtained for a given input excerpt are displayed in this radar chart forming the shadowed area. This facilitates the comprehension by the human user of these scores in their contribution to characterize the possible language level of the input excerpt.

Additionally, the tool offers also several other, secondary metrics, that if necessary, the human operator can resort to in order to refine the elements used to support his judgment. These auxiliary metrics include the frequency, and average and proportion where appropriate, of letters, syllables, words, coordinations and of clauses of three types, viz. simple, passive and subordinate.

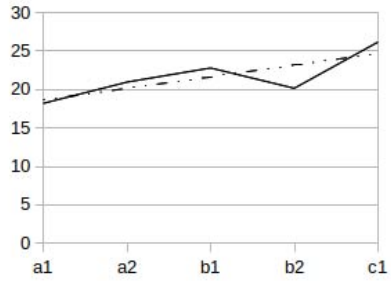
They further include the frequency of occurring words, with the list of words being ordered along the decreasing number of occurrences of tokens per type. And finally these secondary metrics include also counts permitting to assess the lexical density of categories other than nouns, where lexical density is characterized by the number of words occurring per POS.



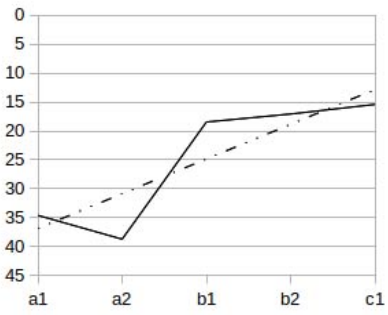
(a) Lexical density in proportion of nouns



(b) Word length in number of syllables per word



(c) Sentence length in nb words per sentence



(d) Flesch index

Fig. 1. The four primary metrics

In its current stage of development, the tool is offered as an online service from a web page with a very lean graphical layout, available at http://lxcefr.di.fc.ul.pt. At the start, that page offers a box to enter text. And when executed, by pushing a button, the service returns that page completed with the scores and elements described above, as in the example presented in Fig. 2.

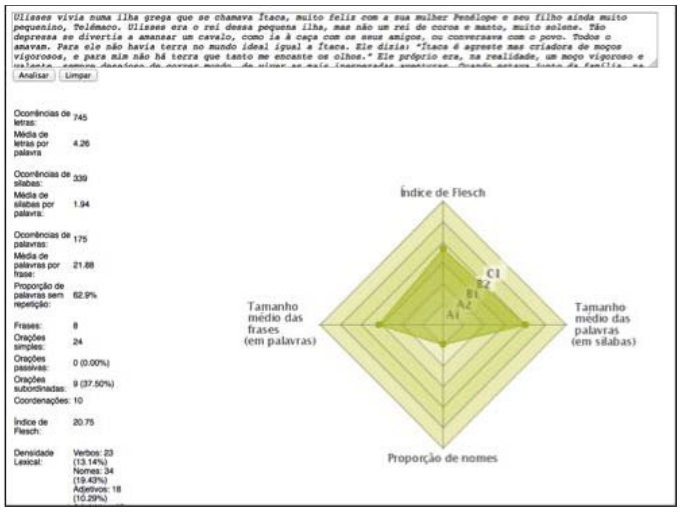


Fig. 2. Example of output by the online service

For any given input, for all metrics, either primary or secondary, absolute scores are provided by the tool and displayed on the left hand side of the pane. Only for the four primary metrics, their scores are also projected into the radar chart, in the right hand side of the pane, thus permitting a ranked reading of its impact in terms of the range of the five language levels considered.

VI. Implementation

Following the design options discussed above, the implementation of this service was based on a number of natural language processing tools whose state of the art performance offers pretty good accuracy: tokenizer, sentence splitter, syllabifier, POS tagger, syntactic constituency parser.

Whenever possible, we resorted to tools previously available with top-level performance and ready to be used. This was possible for all cases except for the syllabifier. Though it is possible to find a couple of publications on the development of syllabifiers for Portuguese (Gouveia *et al.*, 2000; Oliveira *et al.*, 2005), we could not find any syllabifier available, and we eventually developed our own (Rodrigues *et al.*, 2014). We evaluated the syllabifier over the dataset in Porlex (Gomes and Castro, 2003), which included the syllabified representation of over 27,000 lexical entries.

In the remainder cases, we used the tools in the collection of LX tools. The LX-Tokenizer is a two-level tokenization tool, reported as having 99.72% accuracy; in order to isolate sentences, we used LX-Chunker, with a reported 99.94% accuracy (Branco and Silva, 2006). The POS tagger used was the LX-Tagger, based on a maximum entropy approach, and displaying 96.87% of accuracy (Branco and Silva, 2004). The LX-Parser used has a reported performance of 88% F-score (Silva *et al.*, 2010). For some metrics, the NLTK (Bird *et al.*, 2009) chunk package was also used, allowing to identify non-overlapping groups of words with the use of a chunk grammar.

For some of the metrics, it was quite straightforward to obtain the respective scores, namely those involving some counting at the lexical and word levels. For some other metrics, especially those involving some sentential dimension, it was

necessary to implement a number of heuristics that run over the output of some of the above tools. That was the case of the metrics dependent on the counting of the number of simple, subordinate and passive clauses, and the number of coordinations. These heuristics rely on pattern matching over the POS-annotated text and over syntactic trees output by the parser.

The core of the online service was implemented with Python. The user interface uses PHP and communicates with that underlying core through JSON.

VII. EVALUATION

A. Testing the impact of the components of the tool

In order to support the development of the tool, we constructed five test sets used to run regression testing, which are available at http://lxcefr.di.fc.ul.pt . These test sets included 210 input items carefully manually crafted in order to exhibit the full range of cases that were anticipated as key challenges for the different metrics. Their statistics are provided in Table II.

From the error analysis, it became evident that, as expected, the performance of the tool will be faced with two types of failures. On the one hand, there are possible failures induced by errors of the underlying language processing tools (tokenizer, POS tagger, etc.), which are statistically based and have suboptimal accuracy given the state of the art in natural language processing, as detailed above.

On the other hand, there are possible failures due to lack of coverage by the supporting tool given the state of the art of their development. In this respect, a class of grammatical phenomena in point is elliptical sentential constructions, whose occurrence affects particularly the accuracy of the counting for subordinate clauses. But it is worth noting that the counting of subordinate clauses is offered only as a secondary metric, and that the typical frequency of elliptical sentences tends to be very low in most types of text.

In terms of the four primary metrics, displayed in the radar chart, the one that may be more vulnerable to eventual errors of the supporting tools is the proportion of nouns, as it depends on the performance of the POS tagger. Together with the reduced size of the dataset, this circumstance may help to understand the almost flat slope in chart (a) in Fig. 1.

TABLE II. STATISTICS OF THE FIVE DATASETS FOR REGRESSION TESTING

	Items	Tokens	Accuracy
words	11	14	100%
clauses	67	403	98.7%
coordination	27	203	99.8%
passives	21	169	99.8%
subordinates	125	803	63.7%

B. Evaluation of the tool

To evaluate the tool we used the 110 excerpt dataset and performed a 10-fold cross evaluation for each of the four primary metrics, obtaining accuracy values ranging from 9.00% (for word length) to 21.82% (Flesch index), as displayed in the dark grey columns of the chart in Fig. 3.

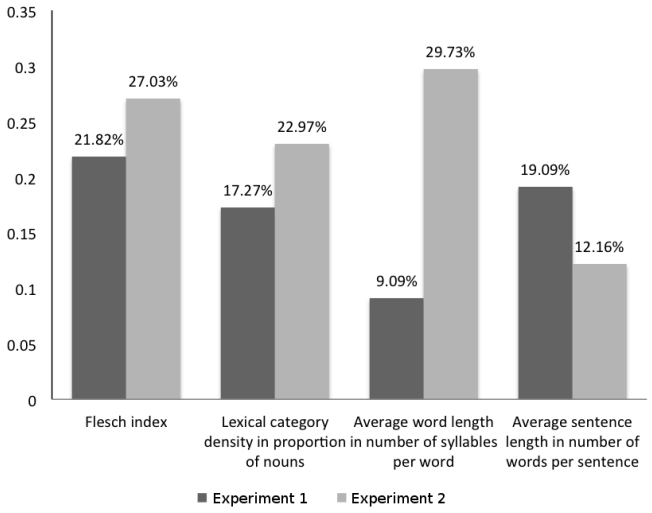


Fig. 3. Evaluation of tool in first and second experiment

C. Assessment of the task

In order to put these scores into perspective, it was important to assess how inherently difficult could the task be in itself, that is how well human annotators executing it could perform.

To that end, the texts in the dataset were untagged of their originally assigned level, and five language instructors were recruited, who are experts trained in selecting and classifying texts according to the relevant five CEFR levels. Each of these experts independently performed the task of classifying each one of the 110 excerpts. This re-annotated dataset permitted to assess the difficulty of the task along two measures: the proportion of texts upon which there is agreement among annotators in their classification; and the inter-annotator agreement (ITA) given by Fleiss' kappa coefficient (Fleiss, 1971; Artstein and Poesio, 2008).

The distribution of the classifications of the annotators per CEFR level is displayed in the chart of Fig. 4.

The excerpts that received unanimous classification by the 5 annotators were 0.90% (only one excerpt); those receiving classification by a majority of at least 4 classifiers were 17.27% (19 excerpts); and there were 67.27% (74 excerpts) receiving the same classification by a majority of at least 3 annotators.

The Fleiss' kappa coefficient value obtained for ITA was 0.13, corresponding to “Slight agreement”, according to Landis and Koch, 1977, which is the second worst in a scale of five levels of agreement, and very distant from the 0.8+ score widely assumed to be the level ensuring reliability of an annotated dataset.

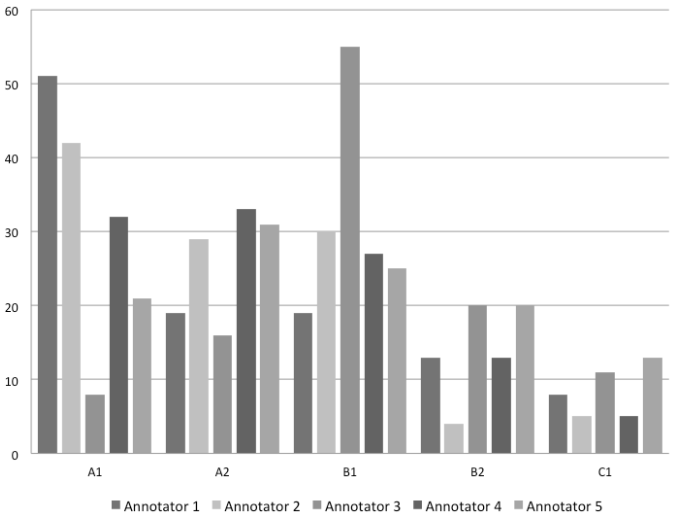


Fig. 4. Distribution of annotators per levels classified

D. Reevaluation of the tool

In order to redeploy our tool on the basis of more reliable empirical basis, from the reannotated dataset, we kept the 74 texts that received its classification by a majority of at least three annotators. With this subdataset with this new and congruent classification by several human experts, the reference scales for the four primary metrics were redone, following the same process as explained above in Section V.

The evaluation of the tool was then also redone, again with a 10-fold cross evaluation of the four primary metrics. The values now obtained after this fine-tuning of the tool with the dataset annotated by multiple experts range from 12.16%, for sentence length, to 29.73%, for word length, with 27.03% and 22.97% for Flesch index and nouns density, respectively, as depicted in the light grey bars of the chart in Fig. 3.

There was a substantial improvement of three of the metrics with respect to their scores obtained in the first evaluation, when the system had been tuned with the dataset of texts used in previous exams, classified by a single instructor.

E. Assessment of the tool by users

The ultimate goal of the tool is to serve as an auxiliary instrument aimed at helping humans in making their judgments about which level assign to a given text. To assess how this goal is being achieved by the current version of the tool, we run an inquiry to collect feedback from users.

The users selected for this assessment were seven expert instructors who had been recruited by the language institute Camões IP to write the actual exams run for language level certification, and who had been using the tool to help themselves in making their judgment concerning candidate excerpts they pondered to include in exams.

They were asked to answer the following two questions, preceded by a contextualization sentence (in Portuguese):

You have been using the text analytics service made available at (http://lxcefr.di.fc.ul.pt).

A. Did this service help you in deciding which language level to assign to each excerpt that was analyzed?

B: Did this service help you in increasing your level of confidence in the classification decisions you have made?

These questions were made available online with the help of Google Docs, permitting the users to answer them in an anonymous fashion. To answer each one of these two questions, the users should select one of the following options, displayed in a dropdown menu.

- nothing (Portuguese: nada)
- little (pouco)
- fairly (razoavelmente)
- quite (bastante)
- a lot (muito)

The statistics of the collected answers are depicted in Fig. 5.

As much as 85% of the users considered that the tool "helps a lot" in their deciding what language level to assign to a given excerpt of interest, and 71% consider that the tool "helps a lot" in increasing their confidence in their classification judgments.

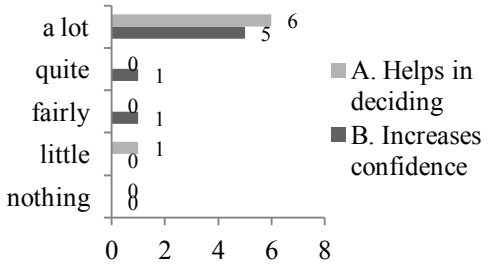


Fig. 5. Assessement of the tool by users

VIII. DISCUSSION

In discussing the results reported in this paper, one should start by noting that the assessment of the task, with a score of 0.13 of ITA, reveals that this is a very difficult task, even for human experts, and hence it represents a quite non-trivial challenge for an automated tool.

On a par with this very low score for ITA, one should consider also that only around 2/3 of the texts (67.2%) that were submitted to be classified by the group of five experts get an identical score by the majority of them. We should thus take this as an upper bound for the performance of the automatic tool, which in its current design setting, happens thus to have no empirical ground to eventually perform better than humans. Given this, we can recall the best accuracy scores obtained by the tool for the four primary metrics, in its second, retrained version, ranging from 29.73% to 12.16% (depicted in Fig. 3). Taking these scores into account, and in particular the score of 27.03% of accuracy obtained with the Flesch index, which combines different linguistic dimensions, one may consider

that the tool is already attaining at least 1/3 of its upper bound, i.e. that its classification performance approximates the performance of humans by at least one third.

In this connection, it should be noted again that the volume of the dataset available may be too small and that there are good reasons to hope that better results may eventually be expected with a larger dataset. In this connection, also worth noting is that, with a larger enough dataset, the superposition of a linear regression upon the distribution of the CEFR levels can be challenged and enhanced, and above all, advanced machine learning methods can be eventually benefited from.

IX. Conclusion

However enticing these results and reasoning may be, one should never lose sight of what was already noted and stressed above: these indexes are meant to be indicators that support human judgment, which should take into account all the other factors that are relevant for the complexity of a text. These metrics are meant to be auxiliary tools for human judgment, not to be stand alone, accurate one-stop predictors.

It should also not go unnoticed that from the first to the second experiment, the accuracy of all four primary metrics improved except one. In the first experiment, each text in the dataset used had been annotated by one single annotator, while in the second experiment, each text taken into account was annotated consistently by the majority of annotators in a group of five. This was the only difference between the two experiments. The substantial improvement in the classification performance of the tool permits thus to hypothesize that this improvement was due to the fact that it was trained over a more consistent dataset.

Clearly this adds again to the observation above, that the task of classifying texts according the CEFR levels of language proficiency is a difficult task, even for experts. But this result also brings to light that this task, and in particular the definition of the criteria for distinguishing the language levels supporting it, may not be well enough specified. On the positive side though, these experiments may also suggest that a possible measure to mitigate this is to resort to the classification of texts by multiple annotators who, by independently annotating texts, happen to agree among each other, especially in the cases where these texts are going to be used in critical situations, including in diploma awarding certification exams.

In any case, it should be noted that the classification of input excerpts according to one of the five relevant language levels, along each one of the four primary metrics (in the radar chart, on the right side of the pane), is just part of the job of this tool. All in all, its ultimate goal is to help humans in their judgment about which level assign to a given text. And for this purpose, it is also important the contribution of all the other metrics, which we termed as secondary (on the left side of the pane). Though their scores are not projected into any ranking scale, they are also important in helping to objectively characterize linguistic dimensions of the excerpt that may be relevant for its eventual classification by the human user.

In this connection, it is worth noting the very positive results obtained when assessing the satisfaction of expert users, with a vast majority of them acknowledging that the tool "helps a lot" both in their eventual classification judgment and their confidence on their judgments. These results are very encouraging especially taking into account that these users have been using what is only a beta version of the tool, which still has quite some room for improvement in the future, as the above remarks in the present section hint at.

Finally, also as future work, it will be important to undertake a more sophisticated usability assessment in order to gain a better understanding of how the tool can possibly be better adjusted to fit the needs of human users, including those that need to use it in their professional duties.

References

[1] Aluísio, Sandra, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. “Readability assessment for text simplification”. In *Proceedings of The 5th Workshop on Innovative Use of NLP for Building Educational Applications*. NAACL-HLT 2010, pp.1-9.

[2] Artstein, Ron and Massimo Poesio, 2008, "Inter-Coder Agreement for Computational Linguistics", *Computational Linguistics*, 34(4), pp.555-596.

[3] Avelãs, Mariana, António Branco, Rosa del Gaudio and Pedro Martins, 2008, "Supporting E-learning with Language Technology for Portuguese", *Lecture Notes in Artificial Intelligence* 5190, Berlin, Springer, pp. 192-201.

[4] Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*, O'Reilly Media Inc.

[5] Branco, António and João Silva, 2004, “Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese”. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa and Raquel Silva (eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC2004), Paris, ELRA, pp.507-510.

[6] Branco, António and João Silva, 2006, "LX-Suite: Shallow Processing Tools for Portuguese", In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics* (EACL2006), Trento, Italy, pp.179-182.

[7] Camões, 2014, https://www.instituto-camoes.pt/epe-inscricoes/certificacao#certificacao_05. Access Date: 1 March, 2014.

[8] Conselho da Europa, 2001, *Quadro Europeu Comum de Referência para as Línguas – Aprendizagem, Ensino, Avaliação*, Alfragide, Edições Asa.

[9] Council of Europe, 2001, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf . Access Date: 3 June, 2014.

[10] Dell’Orletta, Felice, Simonetta Montemagni and Giulia Venturi, 2011, “READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification”. In *Proceedings of the Workshop on “Speech and Language Processing for Assistive Technologies”* (SLPAT 2011), Edinburgh, July 30, 73–83.

[11] DuBay, William, 2004, *The Principles of Readability*, Costa Mesa, Impact Information.

[12] Feng, Lijun, Martin Jansche, Matt Huenerfauth and Noémie Elhadad, 2010, “A comparison of features for automatic readability assessment”. In *Proceedings of the 23rd International Conference on Computational Linguistics* (COLING 2010), pp. 276–284.

[13] Fleiss, Joseph L. 1971. "Measuring Nominal Scale Agreement among many Raters". *Psychological Bulletin*, 76(5):378–382.

[14] Flesch, Rudolf. 1979. *How to write in plain English: A book for lawyers and consumers*. New York: Harper.

[15] Gomes, Inês and São Luís Castro, 2003, ”Porlex, a lexical database in European Portuguese”, *Psychologica*, 32, 91-108.

[16] Gouveia, Paulo, João Teixeira and Diamantino Freitas, 2000, “Divisão Silábica Automática do Texto Escrito e Falado”, *Actas do V PROPOR – Processamento Computacional da Língua Portuguesa Escrita e Falada*, Atibaia, S. Paulo.

[17] Graesser, Arthur C., Danielle S. McNamara, Max M. Louwerse e Zhiqiang Cai. 2004. “Coh-Metrix: Analysis of text on cohesion and language”. *Behavioral Research Methods, Instruments, and Computers*, 36, pp. 193-202.

[18] Instreamia, 2014, http://www.instreamia.com. Access Date: 1 March, 2014.

[19] Landis, J. R., and Koch, G. G., 1977, "The Measurement of Observer Agreement for Categorical Data". *Biometrics*, 33(1), 159-174.

[20] Mitkov, Ruslan (ed.), 2003, *The Oxford Handbook of Computational Linguistics*, Oxoford, Oxford University Press.

[21] Monachesi, Paola, Lothar Lemnitzer and Kiril Simov, 2006, Language “Technology for eLearning”. In W. Nejdl and K. Tochtermann (eds.), *Lecture Notes in Computer Science* 4227, Berlin, Springer, pp. 667-672. Proceedings of the conference on Technology Enhanced Learning (EC-TEL 2006).

[22] Oliveira, Catarina, Lurdes Castro Moutinho, António Teixeira, 2005, “On European Portuguese Automatic Syllabification”, *Proccedings of Interspeech 2005*, pp.2933-2936.

[23] Rodrigues, Erica dos Santos, Cláudia Freitas and Violeta Quental, 2013, “Análise de Inteligibilidade Textual por meio de Ferramentas de Processamento Automático do Português: avaliação da Coleção Literatura para Todos”, *Letras de Hoje*, 48-1, pp.91-99.

[24] Silva, João, António Branco, Sérgio Castro, and Ruben Reis 2010, "Out-of-the-Box Robust Parsing of Portuguese", In *Lecture Notes in Artificial Intelligence*, 6001, pp.86-89, Berlin: Springer. Proceedings of the 9th International Conference on the Computational Processing of Portuguese (PROPOR'10)

[25] Scarton, Caroline E., Daniel M. Almeida, Sandra M. Aluísio. 2009. “Análise da Inteligibilidade de Textos via Ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português”. In *Proceedings of STIL-2009*, São Carlos, Brazil.

[26] Rodrigues, João, Francisco Costa, João Silva and António Branco. Forthcoming. “Automatic Syllabification of Portuguese”, In Actas do XXX Encontro Nacional da Associação Portuguesa de Linguística (ENAPL’14)