First Steps Towards Automatic Multilingual Readability Assessment

Ion Madrazo Azpiazu

Wednesday 2nd March, 2016

1 Introduction

Reading is an important skill in the academic environment, a skill that can be critical for students' educational opportunities and their careers [28]. Giving students texts that suppose an increasing challenge to read during their learning process is essential. Even outside the educational environment, reading plays an important role. It is critical for people to fully comprehend the texts they read, specially when they face medical or legal issues. Understanding a legal or medical document properly, can lead the reader to make a better and more confident decision. However, studies [19, 25, 27] show that even medical documents that are supposed to be suited for average readers, tend to be too specialized and even well-educated adults have trouble understanding. Ensuring that the produced texts are simple enough for people with low reading abilities is imperative.

The readability or complexity of a text, and thus, the audience of it, can be assessed using a readability score. A readability score refers to the degree of ease with which a reader can understand a given text, a score which is usually determined by a readability formula. Historically, teachers have been the main stakeholders of readability formulas,using them to select new materials for their courses and curriculum design. However, lately, readability scores have been known to have more applicability than the ones in academic environments. Automatic text simplification [29, 31], summarization for people with reading difficulties [12], book recommendation [26], literacy assessment [33], or even legal [22] and medical document complexity assess-

ment [19, 25, 27] are only a few examples of applications that take advantage of the comprehension levels generated by existing readability scores.

Traditional formulas such as Flesh [15] became very popular in te late 40's among the educators for manually determining text difficulty. Most of those formulas contained *shallow features* which could be easily adapted to multiple languages and provided a simple way of determining a texts complexity. The multilingualism they provided supposed numerous benefits in environments where the readability of more than one language was needed, i.e. in multicultural environments, education or learning a second language. However, they lacked precision in some cases, such as the one claimed in [9] where nonsense text could be classified as simple to read, just because it contained short and frequently used words. This encouraged researchers to study and develop better and more complex methods of prediction [3, 16], that depended upon natural language processing and machine learning techniques. These new formulas usually continued using the aforementioned shallow features, but added more complex features based on the syntax and semantics of the text. With the addition of new features, the tools became more precise, but more constrained regarding the language adaptability. They started to use increasingly more language-dependent strategies, which made the systems difficult to adapt to be used to estimate readability scores for text in other languages than the one they were designed for, making the multilingualism that was possible in the early stages disappear.

With multilingualism and precision in mind, we propose to develop **MRAS**, a **M**ultilingual **R**eadability **A**ssessment **S**ystem . This tool should both show results comparable to monolingual state of the art systems, and maintain the multilingualism the early tools in the readability field had. For doing so we will explore features and methods used in literature and adapt them to be multilingual. Furthermore, we will develop novel features and analyze the effect each of them has regarding readability. This will also allow us to determine what features increase readability in a text in overall and for specific languages. The created system, will be *open source* and *easily connected* to different application that require readability assessment as a service, potentially permitting the analyze of all sorts of texts, including text snippets, books, websites and even short and unstructured text such as the one found in social media. In doing so, we will produce a system that will adapt itself to the input text language, and use an adequate subset of features for that certain language for giving a prediction, creating, to the best of our knowledge, the first multilingual readability assessment system.

As a byproduct, we will create a leveled dataset for all the tested languages It is important to note, that for practical purposes, the application will only be tested in three different languages: *English*, for state of the art comparison purposes and as reference of germanic languages. *Spanish*, as a reference for romance languages, and *Basque* as an example of a pre-indoeuropean and minority language.

2 Thesis statement

We would like to make an exploration of natural language processing, information retrieval, and social network analysis based features that can aid in the prediction of readability for multiple languages. We would also like compare and analyze different methods for machine learning, in order to see which one can fit best the multilingual readability prediction task.

3 Related work

From the past six decades, different Readability Assessment (RA) systems have been developed with high diversity in terms of both languages and features. Initial readability formulas, such as Flesh [15], Dale-Chall [6], and Gunning FOG [2] made use of **shallow features**, mostly based on ratios of characters, terms, and sentences. These formulas, were simple enough even to be computed manually, providing a simple way of estimating a text's complexity, even if the formulas lacked precision in some cases [9]. This simplicity, however, made them easy to be adapted to estimate readability scores in different languages [30].

In the recent years, the readability formulas have evolved to supervised learning based systems, that show an improved precision by using a combination of old shallow features and new natural language processing based ones. However, incorporating new features has brought a drawback to the area, evidenced by the fact that current systems are too focused in certain languages, and the multilingualism that was possible in the early day is currently lost. Therefore, current state of the art is formed by methods focused on specific languages:

For **English**, the authors of [14] presented a comparison of the common readability features used for English. Some systems [3] were aimed at evalu-

ating text simplification methods making use of elaborated features such as ambiguity among the terms in the texts. Other authors [13] oriented their system for assessing the difficulty level of a text for people with intellectual disabilities by developing features that were intended to detect how well a text was structured. A readability prediction system for finalcial documents [5] was presented, which was based on features such as, presence of active voice or number of hidden verbs. It is also important to mention the currently most used readability assessment formulas Lexile[1] and AR [2], that even if their algorithms are not public, they are widely used among English speaker academic professionals.

In contrast to English, **Spanish** readability assessment has not seen any significant improvement regarding features in the recent years, most of the works being focused in shallow features yet. Several systems [11, 32] have focused their work on text simplification and its evaluation using classical readability formulas. Formulas such as, SSR [30] based on sentence length and number of rare words per sentences or LC and SCI [4] based on density of low frequency words in text.

Compared to other languages **Basque** readability assessment is reduced, to the best of our knowledge, only one system has been developed. Due to the fact that Basque is considered a minority language and shares very little similarities with the most spoken languages, very little research have been done in the area. Therefore, currently, Errexail [18] is the only system created for Basque readability assessment. This system was aimed for text simplification purposes and was developed to predict two different values, simple or complex. The goal of this was to detect which texts needed some simplification and which texts did not. The system makes use of simple features mostly based on ratios of common Natural language processing tags.

Similar to Basque, the literature for **Arabic** readability assessment is reduced too, the authors of [1] developed a readability assessment told based on only two features. The features were based on simple ratios based on sentence,terms and letter counts. Those, features were used with a SVM classifier in order to be able to classify text as simple or complex.

Opposed to previous languages for **Chinese** structure does not look to have such an impact in readability assessment. Therefore, most of the works for Chinese have been focused only on lexical features. The authors of [7]

[1]https://www.lexile.com/

[2]http://www.renaissance.com/products/accelerated-reader/atos-analyzer

developed a RA system based only on lexical metrics based on the TF-Idf measure. However this technique was not topic independent, as once trained for a certain topic the terms were no longer useful for other ones. The authors of [8] developed a system that already tried to solve this issue for Chinese. This system was based on Tf-Idf too and as the authors stated, removing some top scoring words of the Tf-Idf ranking, lead the system to be more independent of the topic.

In contrast to the aforementioned techniques the authors of [10] presented a readability assessment system for **Italian** aimed at assessing readability of sentences. Since the text simplification tool the authors were developing was based on sentences, the authors of this system decided that rather than developing a system for determining text readability, their system would work at sentence level. Therefore, the text simplification tool, would have more information of which sentence needed simplification and which did not. The model generated for sentence level was shown to be generalizable to full text level, by the use of simple averages. The more complex sentences a text have, the more probabilities it have to be complex in overall.

Rather than focusing on the general reader, the authors of [16] developed a readability assessment system for **French**, with the foreign language learners in mind. The objective was to determine which features were more important for a foreign language learner to understand a text. In addition, they provided a metric new to the area called adjacent accuracy that tried to measure systems' performance in a more accurate and relevant way.

Even if the number readability assessment systems that tackle individual languages is high, to the best of our knowledge, no literature exists regarding **multilingual readability** assessment tools, making MRAS a unique system in the area.

4 Proposed Method

We propose to develop MRAS using a supervised learning approach that will rely on knowledge acquired from a leveled corpora. For designing MRAS we will follow the steps illustrated in figure 1 and discussed below.

4.1 Text processing

Different text processing methods have been identified for the development of MRAS. **Freeling NLP** [23,24] is a multilingual natural language processing (NLP) toolkit that supports 11 different languages. This tool solves common NLP tasks such as, tokenization, sentence detection, part of speech tagging or dependency parsing. **WordNet** is a lexical database that takes advantage of semantic relations between terms to build a graph that is very convenient for semantic analysis tasks. **Latent semantic analysis** is also a commonly used strategy for semantic analysis, which takes advantage of concurrences among terms for determining similarities between them. All those tools and others that we will discover in the process of development, will form the text processing step of MRAS.

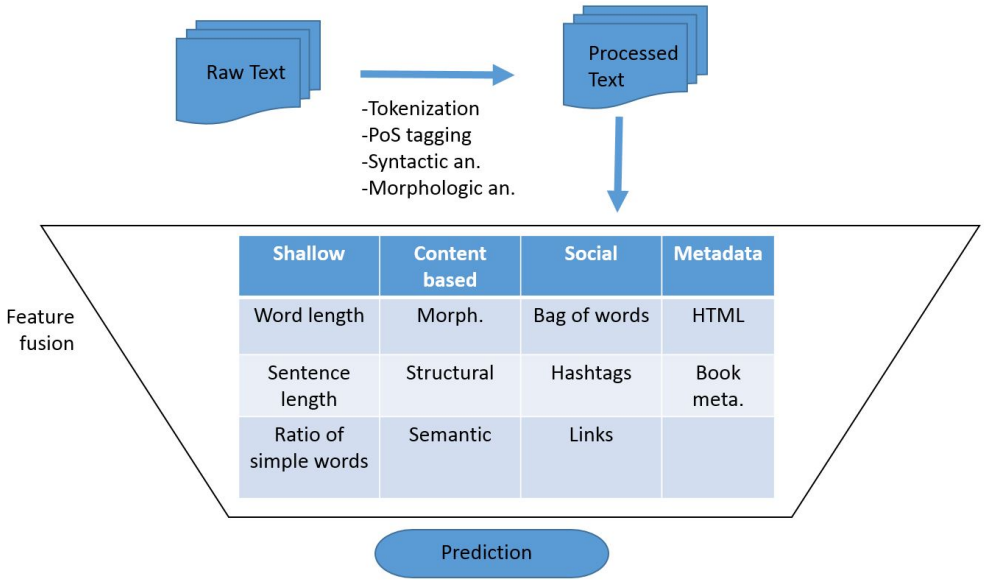

Figure 1: Description of MRAS

4.2 Feature extraction

The exploration of features will be one of the main tasks of this thesis. MRAS should be able to extract a wide range of features that satisfy the needs of

each language it will tackle. A general description of the types of features that we expect to incorporate in MRAS can be seen below.

Shallow features Shallow features [2, 6, 15] have historically shown to be of good use when prediction readability. Therefore, they will be incorporated into MRAS and used as a baseline for improvement. Sentence length, word length, or ratio of simple terms are examples of the features that will be included among this subset.

Morphological features Morphology analyses how word-forms are formed from its root. Even if this aspect is not relevant in some languages such as English, it has been shown to be strongly correlated with the readability in some languages such as Basque [18]. Different morphological phenomena will be analyzed in order to create features in this subset.

Structural features Structural features are the ones that describe how a text is organized. They can both describe structure within the sentence (syntactical structure) or structure between sentences (pragmatical structure). Depth of the syntactic tree or ratios of different types of connectors between sentences are some examples of the features that are going to be explored under this subset.

Semantic features Semantic features go beyond the tokens and structure of the text in order to analyse the concepts laying on it. This, permits to create an abstraction level that leave behind the dependence other features have respect to the text. Features such as, concept density or concept followability are some examples of the features that will be analyzed under this subset.

Social features Readability assessment can be used in more than just plain text. Internet is evolving into a new social era and so are text resources too. Increasingly more resources contain social information such as hashtags, mentions or links, an information that is usually ignored by readability formulas. We would like to perform a preliminary research in order to see how the aforementioned information can be used for readability prediction.

4.3 Prediction

The aforementioned features need to be fusioned in order to make a prediction. For this, would like to explore different prediction strategies for developing MRAS. The problem of assessing readability can be seen as a classification problem where a discrete categorical class needs to be predicted. Therefore, we would like to explore different **classification** algorithms, such as bayesian networks or support vector machines. Me subrayaste "machines", no entiendo el porque The readability assessment task can also be seen as a **regression** problem, given that the class contains an inherent order on it. Therefore, we will also like test different regression algorithms. Finally, we would also like to take an **hybrid** approach by using classification algorithms that take order in the class into account, such as the ordinal classification approach presented in [17].

5 Evaluation

Even if MRAS is designed to work in many more languages, for practical purposes the evaluation will only be carried in three languages, that we think can faithfully represent the diversity of existing languages. For this purpose, we have chosen a germanic, a romance, and a pre-indioeuropean language, i.e. English, Spanish, and Basque respectively.

5.1 Datasets

The ideal dataset for developing MRAS would be a multilingual leveled dataset that would contain the exact same documents written in different languages. However, to the best of our knowledge, such a dataset does not currently exist. Therefore, we have identified, various sets of leveled documents for each individual language that can suit MRAS' needs and can be used for evaluation purposes. Those can be seen in table 1.

5.2 Metrics

The performance of MRAS will be evaluated by means of (1) common classification evaluation methods, such as accuracy and error type rates, (2) regression evaluation methods such as RMSE (Root mean square error) and

	Dataset	Description
English	Lexile	Contains book titles associated with its readability level
	Stantarized tests	Tests for English level, they contain various texts per test
	Other	News for kids, exercises for learning English
Spanish	Lexile	Contains book titles associated with its readability level
	Learning resources	Various exercises for learning Spanish
Basque	Learning resources	Various exercises for learning Basque
Multilingual	Paraller corpus	Contains same texts translated into two languages

Table 1: Data resources identified for MRAS development

(3) methods used in the readability assessment area, such as adjacent accuracy [16].

5.3 Overall assessment

The study and performance analysis of this thesis will aim at answering the following questions:

- Which learning model performs better for MRAS? Which feature subset?
- Which features add more value in terms of predicting readability? Do they add same value for each language?
- How does MRAS perform compared to baseline shallow feature based formulas? and compared to state of the art systems?
- Would MRAS give same prediction for the a text that is translated manually into another language? and for a text that is automatically translated?
- How efficiently can MRAS predict the readability of a language for which it has not learned? If we train MRAS for two languages can we use it to predict the readability of a text in a third one?
- If we have a really small dataset for one language, would adding more data from another language improve the prediction results of the first one?

6 Proposed schedule

- Datasets created
- Shallow features
- Content based features
- Social features
- Prediction algorithm
- Evaluation
- Any conference?
- Documentation
- Presentation

References

[1] A. A. Al-Ajlan, H. S. Al-Khalifa, and A. Al-Salman. Towards the development of an automatic readability measurements for arabic language. In *Digital Information Management, 2008. ICDIM 2008. Third International Conference on*, pages 506–511. IEEE, 2008.

[2] J. Albright, C. de Guzman, P. Acebo, D. Paiva, M. Faulkner, and J. Swanson. Readability of patient education materials: implications for clinical practice. *Applied Nursing Research*, 9(3):139–143, 1996.

[3] S. Aluisio, L. Specia, C. Gasperin, and C. Scarton. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. Association for Computational Linguistics, 2010.

[4] A. Anula. Tipos de textos, complejidad lingüstica y facilicitación lectora. In *Actas del Sexto Congreso de Hispanistas de Asia*, pages 45–61, 2007.

[5] S. B. Bonsall, A. J. Leone, and B. P. Miller. A plain english measure of financial reporting readability. *Available at SSRN 2560644*, 2015.

[6] J. S. Chall and E. Dale. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books, 1995.

[7] Y.-H. Chen, Y.-H. Tsai, and Y.-T. Chen. Chinese readability assessment using tf-idf and svm. In *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*, volume 2, pages 705–710. IEEE, 2011.

[8] K. Collins-Thompson and J. P. Callan. A language modeling approach to predicting reading difficulty. In *HLT-NAACL*, pages 193–200, 2004.

[9] A. Davison and R. N. Kantor. On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading research quarterly*, pages 187–209, 1982.

[10] F. Dell'Orletta, S. Montemagni, and G. Venturi. Read-it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, pages 73–83. Association for Computational Linguistics, 2011.

[11] B. Drndarević, S. Štajner, S. Bott, S. Bautista, and H. Saggion. Automatic text simplification in spanish: a comparative evaluation of complementing modules. In *Computational Linguistics and Intelligent Text Processing*, pages 488–500. Springer, 2013.

[12] L. Feng. Automatic readability assessment for people with intellectual disabilities. *ACM SIGACCESS Accessibility and Computing*, (93):84–91, 2009.

[13] L. Feng. Automatic readability assessment for people with intellectual disabilities. *ACM SIGACCESS Accessibility and Computing*, (93):84–91, 2009.

[14] L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284. Association for Computational Linguistics, 2010.

[15] R. Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948.

[16] T. François and C. Fairon. An ai readability formula for french as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477. Association for Computational Linguistics, 2012.

[17] E. Frank and M. Hall. *A simple approach to ordinal classification*. Springer, 2001.

[18] I. Gonzalez-Dios, M. J. Aranzabe, A. D. de Ilarraza, and H. Salaberri. Simple or complex? assessing the readability of basque texts. In *Proceedings of COLING*, volume 2014, 2014.

[19] A. M. Ibrahim, C. R. Vargas, P. G. Koolen, D. J. Chuang, S. J. Lin, and B. T. Lee. Readability of online patient resources for melanoma. *Melanoma research*, 26(1):58–65, 2016.

[20] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

[21] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[22] J. R. Ogloff and R. K. Otto. Are research participants truly informed? readability of informed consent forms used in research. *Ethics & Behavior*, 1(4):239–252, 1991.

[23] L. Padr, M. Collado, S. Reese, M. Lloberes, and I. Castelln. Freeling 2.1: Five years of open-source language processing tools. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC'10)*, La Valletta, Malta, May 2010.

[24] L. Padr and E. Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May 2012. ELRA.

[25] C. R. Patel, S. Sanghvi, D. V. Cherla, S. Baredes, and J. A. Eloy. Readability assessment of internet-based patient education materials related to parathyroid surgery. *Annals of Otology, Rhinology & Laryngology*, page 0003489414567938, 2015.

[26] M. S. Pera and Y.-K. Ng. Automating readers' advisory to make book recommendations for k-12 readers. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 9–16. ACM, 2014.

[27] J. Petkovic, J. Epstein, R. Buchbinder, V. Welch, T. Rader, A. Lyddiatt, R. Clerehan, R. Christensen, A. Boonen, N. Goel, et al. Toward ensuring health equity: Readability and cultural equivalence of omeract patient-reported outcome measures. *The Journal of rheumatology*, 42(12):2448–2459, 2015.

[28] R. D. Robinson, M. C. McKenna, and J. M. Wedman. Issues and trends in literacy education. 2000.

[29] H. Saggion, S. Štajner, S. Bott, S. Mille, L. Rello, and B. Drndarevic. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):14, 2015.

[30] S. Spaulding. A spanish readability formula. *The Modern Language Journal*, 40(8):433–441, 1956.

[31] S. Štajner, R. Mitkov, and G. C. Pastor. Simple or not simple? a readability question. In *Language Production, Cognition, and the Lexicon*, pages 379–398. Springer, 2015.

[32] S. Štajner and H. Saggion. Readability indices for automatic evaluation of text simplification systems: A feasibility study for spanish. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013), Nagoya, Japan*, pages 374–382, 2013.

[33] B. D. Weiss, M. Z. Mays, W. Martz, K. M. Castro, D. A. DeWalt, M. P. Pignone, J. Mockbee, and F. A. Hale. Quick assessment of literacy in primary care: the newest vital sign. *The Annals of Family Medicine*, 3(6):514–522, 2005.