First Steps Towards Automatic Multilingual Readability Assessment

Ion Madrazo Azpiazu

Thursday 10th March, 2016

1 Introduction

Reading is an important skill in the academic environment, a competence that can be critical for students' educational opportunities and their careers [40]. Studies [33] show that reading for learning takes place best, when the reader comprehends a 75% of the text. This supposes an appropriate balance, that allows the reader to positively understand the text, while also finding challenges in the reading process that will motivate him to improve his skills. Outside educational environment, reading takes place for comprehension rather that for learning. Therefore, in this context, it is critical to provide people with texts that they can fully comprehend. E.g. understanding a legal or medical document properly, can lead the reader to make a better and more confident decision. However, studies [32, 37, 39] 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 them. Every reader's skills are different and the complexity of the texts they need to face depends also upon the context. Therefore, providing institutions and average users with tools that permit to assess whether a text is adequate for a user is imperative. That is where *Readability Assessment (RA)* tools take place, providing a way to determine the degree of ease with which a reader can understand a given text, i.e. the *Readability Score (RS)*.

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 [41, 43], summarization for people with reading difficulties [25], book recommendation [38], literacy assessment [45], or even legal [34] and medical document complexity assessment [32,37,39] are only a few examples of applications that take advantage of the comprehension levels generated by existing readability scores. Even in commercial environments, book publishers require from professional linguistic services in order to tag their publications with a readability score, a task that would similarly be achieved by an automatic tool.

Traditional formulas, such as Flesh [28] became very popular in the late 40's among educators for manually determining text difficulty. Most of those formulas relied on *shallow features* to estimate text difficulty, which could be easily adapted to multiple languages and provided a simple way of determining text complexity. The multilingualism they provided supposed numerous benefits in environments where the readability of more than one language was needed, i.e. book translation or learning a second language. However, the traditional formulas lacked precision in some cases, such as the case described in [22] 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, that depended upon natural language processing and machine learning techniques [13, 29]. These new formulas usually continued taking advantage of 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 their language adaptability [15,27]. In fact, they used 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. As a result, the multilingualism that was possible in the early stages disappeared.

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 conceive novel features and analyze the effect each of them has regarding readability. This will also allow us to determine which features determine readability in a text in overall and for

specific languages. MRAS will be *open source* and *easily connected* to different applications that require readability assessment as a service, potentially permitting the analysis 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 the corresponding language for readability prediction, creating, to the best of our knowledge, the first multilingual readability assessment system.

As a byproduct, we will create a leveled dataset with readability labeled documents for different languages. In addition, we will perform an in-depth assessment of existing strategies for readability prediction.

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

Make an exploration and desing natural language processing, information retrieval, and social network analysis based features to aid in the prediction of readability for multiple languages. 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 RA systems have been developed with high diversity in terms of both languages and features. Initial readability formulas, such as Flesh [28], Dale-Chall [17], and Gunning FOG [12] 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 [22]. This simplicity, however, made them easy to be adapted to estimate readability scores in different languages [42].

In the recent years, readability formulas have evolved to supervised learn-

ing based systems, that show an improved precision by using a combination of traditional shallow features and new natural language processing based ones i.e. based on language aspects, such as syntax or semantics. 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 in [27] presented a comparison of the common readability features used for English. Some works [13] were aimed at creating a RA system for evaluating automatic text simplification quality, making use of elaborated features such as ambiguity among the terms in the texts. Other authors [26] 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 [16] 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 commercial RA formulas Lexile[1] and AR [2], which are widely used among English speaker academic professionals. Even if their algorithms are not public, they are known to use shallow features showing how common terms of the text are and how long sentences are in average [33].

In contrast to English, **Spanish** RA has not seen any significant improvement regarding features in recent years, as most of the works are still based in shallow features. Several systems [24,44] focused on combining traditional formulas such as, SSR [42] based on sentence length and number of rare words per sentences or LC and SCI [14] based on density of low frequency words in text.

Compared to other languages **Basque** RA is reduced, to the best of our knowledge, to only one system. Due to the fact that Basque is considered a minority language and shares very little similarity with most spoken languages, very limited research have been done in the area. So far, Errexail [31] was the only system created for Basque RA. This system was aimed for text simplification purposes and was developed to predict two different readability values, simple or complex. The goal of this was to detect which texts needed some simplification and which texts did not. The system used features mostly based on ratios of common Natural language processing tags.

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

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

Similar to Basque, the literature for **Arabic** RA is reduced too, the authors of [11] developed a RA 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 Support Vector Machine classifier in order to be able to classify text as simple or complex.

Opposed to previous languages, structural features do not look to have such an impact for **Chinese** RA. Therefore, most of the works for Chinese have been focused only on lexical features. The system introduced in [18] was only based 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. Another system [19] 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 [23] presented a RA system for **Italian** aimed at assessing readability at sentence level. The developed system was intended to be part of a simplification tool the authors were also developing, which required of a RA system that worked on sentences rather than on full texts.

Rather than focusing on the general reader, the authors of [29] developed a RA 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. They tested lexical, syntactical and semantic features, and showed that semantic ones performed poorly in their case.

Even if the number of RA 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.

Even if the number of RA systems that tackle individual languages is high, they are usually focused on a specific set of features and materials they can analyze. In addition, to best of our knowledge, none the RA systems presented are **multilingual**. MRAS will be focused on adressing all the issued mentioned above. Not only it will be multilingual, but will implement a comprehensive list of features and add new ones to it. Furthermore, the algorithm of MRAS will be general enough to potentially be able to handle all sorts of reading material. All those features will make 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. In designing MRAS we will follow the steps illustrated in Figure 1 and discussed below.

4.1 Text processing

As the main focus of MRAS is to analyze text, we have identified different text processing methods and tools that will be used in its development. **Freeling NLP** [35,36] 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, along with others, that we will incorporate during the research process, will be used part of the text processing step of MRAS.

4.2 Feature extraction

Exploring 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 categories of features that we expect to incorporate in MRAS is presented as follows:

Shallow features [12,17,28] 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 category.

Morphological features capture how terms are formed from their root. Even if this aspect is not relevant in some languages such as English, it has been shown to be a strong predictor for readability in some languages such as Basque [31]. Different morphological phenomena will be analyzed in order

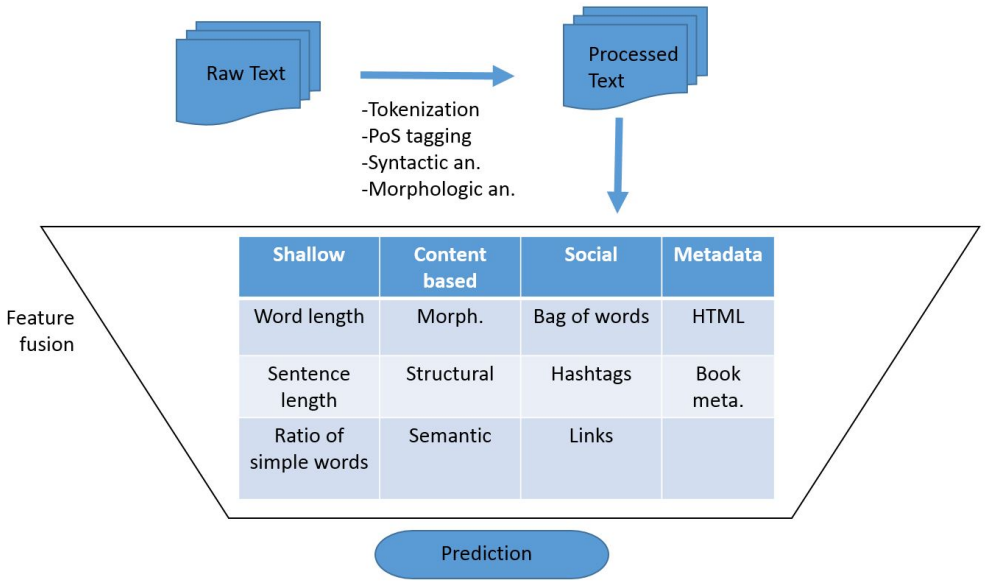

Figure 1: Description of MRAS

to create features in this category.

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 category.

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 follow-ability are some examples of the features that will be analyzed under this category.

RA can be used in more than just plain text. Internet is evolving into a new social era and so are text resources. Increasingly more resources contain **social information**, such as hashtags, mentions or links, this type of information usually ignored by readability formulas. We would like to investigate

how the aforementioned information can be used for readability prediction.

4.3 Prediction

Individually each of the aforementioned features can only provide a raft estimate of the readability of a text. However, considering these features in tandem can lead to a more accurate an robust readability assessment. Consecutively, we will analyze different fusion strategies for 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 [20] for readability level prediction. The RA 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 [30].

5 Evaluation

Even if MRAS is designed to be language independent, for practical purposes the evaluation will only be conducted 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, as well as human judgments, in terms of readability scores for each document. However, to the best of our knowledge, such a dataset does not currently exist. Consequently, we have identified various sets of leveled documents for each individual language that can suit MRAS' needs and can be used for evaluation purposes. Details on the datasets considered for evaluation purposes can be seen in Table 1.

	Dataset	Description
English	Lexile [1]	Contains book titles associated with its readability level
	Stantarized tests [2,3]	Tests for English level, they contain various texts per test
	Other [4–6]	News for kids, exercises for learning English
Spanish	Lexile [1]	Contains book titles associated with its readability level
	Learning resources [7–9]	Various exercises for learning Spanish
Basque	Learning resources [1]	Various exercises for learning Basque
Multilingual	Parallel corpus [10]	Contains same texts translated into Spanish and English

Table 1: Data resources identified for MRAS development

5.2 Metrics

The performance of MRAS will be evaluated by means of (1) common classification evaluation methods, such as absolute error [21], (2) regression evaluation methods such as MSE (Mean square error) [21] and (3) methods used in the readability assessment area, such as adjacent accuracy [29].

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

For conducting the research proposed in this manuscript in a timely manner, we define the milestones shown in Table 2.

Date	**Milestone**
April 2016	Gather existing datasets for design and development
May 2016	Feature Exploration
June 2016	Feature fusion
July 2016	Experiments
July 2016	Thesis draft
August 2016	Defense

Table 2: Proposed schedule

References

[1] http://www.ikasbil.eus.

[2] http://www.flo-joe.co.uk/exams.htm.

[3] http://learnenglishteens.britishcouncil.org/content?field_language_level_tid=50&field_section_tid=1129&field_topics_tid=&language=en.

[4] https://www.readinga-z.com/books/leveled-books/.

[5] http://www.breakingnewsenglish.com/news-for-kids.html.

[6] http://www.newsinlevels.com/.

[7] http://cvc.cervantes.es/aula/lecturas/.

[8] http://aprenderespanol.org/lecturas/lecturas-ejercicios.html.

[9] http://www-k6.thinkcentral.com/content/hsp/reading/Senderos/na/common/online_senderos_libros_graduables_para_lectores/senderos_SE/launch.html.

[10] http://albalearning.com/audiolibros/textosparalelos.html.

[11] 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.

[12] 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.

[13] 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.

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

[15] R. G. Benjamin. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1):63–88, 2012.

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

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

[18] 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.

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

[20] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[21] W. B. Croft, D. Metzler, and T. Strohman. *Search engines: Information retrieval in practice*, volume 283. Addison-Wesley Reading, 2010.

[22] 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.

[23] 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.

[24] 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.

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

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

[27] 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.

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

[29] 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.

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

[31] 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.

[32] 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.

[33] C. Lennon and H. Burdick. The lexile framework as an approach for reading measurement and success. *electronic publication on www. lexile. com*, 2004.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

[41] 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.

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

[43] 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.

[44] 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.

[45] 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.