Automatic Text Classification using Readability Levels in Galician and Spanish

Sandra Rodríguez Rey

Centro Singular de Investigación en Tecnoloxías Intelixentes (CITIUS), Universidade de Santiago de Compostela, 15782, Santiago de Compostela, Spain

Abstract

A text can be more or less complex to read depending on various parameters, such as sentence length, vocabulary variety, or the presence or absence of certain syntactic structures and linguistic phenomena. This concept of the level of reading complexity is called readability. This research aims to study automatic text classification in Galician and Spanish, focusing on the creation of corpora and the development of an automatic text classifier based on readability levels for Galician. First, the state of the art in text readability and automatic text classification is reviewed. Then, corpora of Galician and Spanish texts are compiled and classified by readability levels, and linguistic phenomena of different complexity levels are studied. Finally, different computational strategies for automatic text classification are evaluated.

Keywords

Text classification, text complexity, readability, Automatic Readability Assessment

1. Introduction

Readability has long been studied, from the 19th century until now [1]. Knowing the degree of reading complexity of a text is important in several domains, such as language learning or automatic readability assessment [2]. Readability formulas have been the traditional method of calculating reading complexity for decades. These formulas extract metrics such as word and sentence length [1]. Nowadays, the most common methods to measure the readability level of a text are based on linguistic features or on the use of deep learning models [3]. The latter models are trained on corpora labeled based on readability levels.

Automatic text classifiers based on complexity levels are useful tools for determining the readability level of a text. FABRA [4] is a good example of such a tool. To train automatic classifiers, high-quality corpora labeled by complexity levels are needed. Some corpora are available for Spanish, such as Coh-Metrix-Esp [5], Newsela [6], CAES [7], or Simplext [8]. Nevertheless, these corpora contain texts from only three domains (literature, journalism, and education) and are designed for text simplification or language teaching as a foreign language. To the best of our knowledge, there are no corpora or automatic classifiers available for Galician.

The main objective of this PhD project is the investigation of automatic methods for the assessment of the readability of texts for adults in Galician and Spanish. To achieve this, three specific objectives have been defined. First, to obtain Galician and Spanish corpora for testing

Doctoral Symposium on Natural Language Processing 2024

sandrarodriguez.rey@usc.es (S. Rodríguez Rey)
0000-0002-4864-7255 (S. Rodríguez Rey)

automatic text classifiers. Second, to test the performance of available automatic text classifiers by readability levels for Galician and Spanish. And thirdly, to explore data augmentation and transfer learning strategies to develop new models for Galician.

This research project aims to contribute to the field by answering three research questions (RQs). RQ1: Are reading complexity descriptors similar for Galician, Spanish and other languages? The hypothesis for this research question is that they are similar, but that each language has specific linguistic phenomena that affect text complexity. RQ2: Is it possible to adapt reliable text classifiers designed for other languages with minor modifications? The hypothesis is that it depends on the size of the corpus and the linguistic resources available, so the situation is different for Galician and Spanish. RQ3: Since Galician is considered a language with few resources, is it possible to use cross-linguistic strategies? The hypothesis is that it is possible by using multilingual models and adapting resources from other languages to Galician.

This paper gives a description of the research project and some results. First, some background on the topic is given and some related work is highlighted. Then, a description including research questions, hypotheses and objectives is provided, followed by the methods and techniques used. Finally, a short introduction to the research work carried out during part of the research, and some discussion questions are presented. Regarding the first results of this research project, two corpora for automatic text classifier testing are presented: a Spanish and a Galician corpus of texts for adult population labeled by levels of complexity.

2. Background and related work

Readability is the ease or difficulty with which a text can be read and understood [1]. The first studies on the characteristics that define the level of complexity of a text to be understood date back to the 19th century and took place in the United States and Russia. These studies refer to the lexicon and sentence length [1]. Vajjala [2] refers to Thorndike, Lively and Pressey, and Vogel and Wash-burne as the first studies on how to measure the level of difficulty of reading a text in the 1920s decade.

The traditional calculation method is the application of readability formulas. These formulas measure some superficial characteristics of the text [3], such as sentence length, number of syllables in words or readers' knowledge or unfamiliarity with the lexicon and extract a value indicating the reading complexity of the text [1]. An example is the Flesh-Lincaid metric (1975), which calculates mainly the number of words per sentence and the number of syllables per word [3]. Following Campos [1], Du Bay states that other well-known formulas are the Flesch's Reading Ease Score (RES), the Dale-Chall, the SMOG, the Gunning FOG test or the Fry's Chart.

More recently, various methods and techniques have been used to automatically evaluate the readability of a text. The most common ones are based on linguistic features or deep learning models. The former focuses on word and sentence length, syntactic complexity, or the percentage of occurrence of words included in various types of word lists [2]. There are two types of linguistic feature models that can be highlighted: from the statistical computation of features in the text, or from trained machine learning models. Neural networks and classification algorithms are some of the deep learning methods used [3]. These models are trained on corpora that are labeled based on readability levels.

In the last 10 years, however, significant progress has been made using sophisticated NLP techniques, such as automatic parsing and statistical language modeling [3]. This makes it possible to evaluate a wide list of factors that affect the readability of a text. An example is FABRA, a tool developed for French, which measures several factors, such as word and sentence length, lexical diversity, orthographic neighborhood, lexical frequency, syntactic dependency, syntactic coherence, the use of anaphoric elements, etc. [4]. However, to the best of our knowledge, there are no resources or tools available for our target languages that aim to classify texts of different genres according to readability levels.

3. Justification

Information about the readability level of texts is useful in diverse fields: language learning, automatic readability assessment, content creation, accessibility, etc. [2]. Thus, tasks such as appropriate reading materials selection or text adaptation to make texts clearer and more accessible can be facilitated.

To obtain text readability information, automatic text classifiers are an optimal tool. To develop automatic text classifiers, high-quality corpora are needed. Concerning automatic text classification models, we are not aware of the existence of this type of tool for Galician.

Several corpora on text complexity already exist for various languages: Weekly Reader [9], WeeBit [10], (CLEAR) [11]) or OneStopEnglish [12] for English; FLE-CORP [4], FLM-CORP [4] or FSW [13] for French; READ-IT [3] or CELI [14] for Italian; a dataset from Instituto Camões [15] for Portuguese; Slovenian SB for Slovenian [16]; LBSPC [17] for Basque, or VikiWiki [18] for Basque and Catalan.

For Galician, we are not aware of any corpus of this type that is available. For Spanish, some existing corpora are Coh-Metrix-Esp [5], Newsela [6], CAES [7], Simplext [8], and kwiziq and HablaCultura corpora [19]. Although more Spanish corpora exist, they are unavailable or they have data protection licenses that prevent their use [19]. Spanish available resources include a limited variety of textual genres and domains: literature, journalism, and teaching. Moreover, some of them are designed for text simplification (Newsela or Simplext), and some others for language teaching as a foreign language (CAES, kwiziq, or HablaCultura).

Therefore, this research work aims to contribute to the field by creating new corpora for Galician and Spanish and by exploring reliable automatic text classifiers for Galician texts.

4. Research description

Text readability, understood as a predictor of text complexity for comprehension, affects reading performance [1]. Automatic readability assessment tools make it possible to know the complexity level a text represents for a reader and even certain characteristics that make texts more complex to understand. This information can be valuable for different purposes, such as selecting and simplifying texts for foreign language learners. Despite the number of Spanish speakers, the development of this type of tool for Spanish is scarce [19]. To the best of our knowledge, Galician, which is considered a low-resource language, does not have a linguistic tool with this functionality available.

The main objective of this thesis is to study automatic methods to evaluate the readability of texts for the adult population in Galician and Spanish. To achieve it, three specific objectives have been defined:

- To obtain Galician and Spanish corpora that represent the wide variety of text genres and topics an adult reader might encounter in his or her lifetime classified using readability levels for testing automatic text classifiers.
- To evaluate the performance of several automatic text classifiers by readability levels for Galician and Spanish.
- To study data augmentation strategies by adapting existing resources from other languages to Galician and transfer learning methods using multilingual models to develop reliable text classifiers for Galician.

The following research questions (RQs) and corresponding hypotheses (Hs) are formulated in this thesis:

- RQ1: Are readability complexity descriptors similar for Galician, Spanish and related languages?
  H1: Complexity descriptors of text readability are similar for Galician and Spanish, and also for related languages such as Portuguese or French. However, each language has specific linguistic phenomena that affect the complexity of the text and must be taken into account when determining the readability complexity descriptors. For Galician, some examples can be the use of the dative of interest or the solidarity pronouns, some periphrases (such as the construction "dar" + participle) or the inflected infinitive (although this last one also exists in Portuguese).
- RQ2: Is it possible to quickly adapt reliable text classifiers designed for other languages to Galician and Spanish?
  H2: This depends on the size of the corpus and the linguistic resources available. For Spanish, it may be possible, since there are resources available for measuring text complexity (for example, word lists of concrete and abstract nouns, age of acquisition or graded lexicons) and a corpus has already been created. For Galician it may be possible, but the classifier will probably have a lower performance because the computational resources for Galician are limited.
- RQ3: Since Galician is considered a language with few resources, is it possible to use cross-linguistic strategies to obtain new data and develop an automatic text classifier?
  H3: This is possible by adapting resources from other languages (e.g., Spanish, Portuguese and French) to Galician using techniques such as transliteration and machine translation and by using multilingual models and transfer learning methods.

5. Methodology

To meet the aforementioned objectives, the following methodology is proposed.

First, a review of the state of the art in text readability and automatic text classification according to text complexity levels is carried out, including general parameters for different languages and specific linguistic phenomena for Galician and Spanish. Second, a corpus of about 2000 Spanish texts classified by complexity levels will be created. Thirdly, a similar corpus of about 400 texts in Galician will be compiled. Then, available Transformer models for Galician and Spanish (both monolingual and multilingual) will be evaluated on text classification. Depending on the performance, additional data for Galician may be created by exploring various data augmentation methods. Subsequently, the evaluation of the resulting automatic text classifiers for Galician will be performed.

The main research techniques will be documentary and experimental. For the experimental techniques, both quantitative and qualitative results will be obtained. Regarding the development of linguistic tools for Galician texts, since Galician is considered a language with few resources, the performance of classifiers designed for related languages will be studied [20]. Spanish, Portuguese and French tools will be the main options to study, as these are the languages involved in the project the author is engaged in and similar tools are being developed. Both quantitative and qualitative analyses will be carried out using experimental techniques.

After obtaining these results, we can explore methods such as synthetic data augmentation using generative models, machine translation, or other transformations from related languages [21] if needed. Documentary techniques and questionnaires addressed to language experts and students will be used to determine the characteristics that influence text readability in Spanish and Galician. On the one hand, literature on readability and similar concepts, such as reading comprehension and text simplification for easy reading will be reviewed. On the other hand, specific linguistic aspects affecting readability in Galician, Spanish, Portuguese and French will be studied. Since, to the best of our knowledge, there are no studies on Galician linguistic phenomena that affect readability, specific aspects of Galician-related languages that may also occur in Galician will be analyzed.

Different types of automatic text classification models will be tested, from classical machine learning models like Decision Stress and SVM, as well as the fine-tuning of language models (such as BERT) for classification tasks using open-access libraries, such as Transformers by HuggingFace. Both monolingual and multilingual models will be explored.

6. Results

The research progress made includes the creation of two corpora based on readability levels. Both resources include texts from a wide variety of genres and three levels of complexity defined by experts.

The first corpus is being developed within the iRead4Skills project (Intelligent Reading Improvement System for Fundamental and Transversal Skills Development), an ongoing European project involving researchers from different institutions, such as the NOVA University of Lisbon, the INESC-ID Research Center (Lisbon), the Catholic University of Louvain (Belgium), the University of Santiago de Compostela, the UAB (Universitat Autònoma de Barcelona) and the Luxembourg Institute of Socio-Economic Research. This project aims to improve the reading skills of the adult population with low literacy levels by creating an intelligent system that analyzes text complexity and provides appropriate reading materials, thus facilitating their

Table 1
Spanish corpus: Texts distribution by categories and levels of complexity

Categories	L1	L2	L3	L4
personal communication	149	63	46	6
institutional communication	29	55	92	61
social media	24	84	143	79
commercial communication	111	82	85	20
didactic book	29	37	52	16
fiction book	199	73	86	11
non-fiction book	63	150	167	57
academic	6	27	106	66
political	6	25	44	17
legal	8	17	31	21
religious	39	47	37	0
total	660	660	889	354

access to information and culture[1].

This Spanish corpus, to which the author of this article is one of the two main contributors, is part of a multilingual dataset that includes three corpora with similar characteristics in three languages: Spanish, Portuguese and French [22]. This corpus contains 2563 Spanish texts classified into three levels of complexity: level 1 (very easy), level 2 (easy), and level 3 (plain). A fourth level of more complex texts is also included, although it is not a consolidated level. This fourth level is intended to represent the type of text in terms of complexity that should not be considered level 3. The texts are also classified by text domains, genres, and subgenres, as shown in Table 1. These categories are intended to represent the most common textual and thematic genres that an adult reader might encounter, focusing on the types of texts that are of most interest to an adult reader.

The second resource to be presented is a Galician corpus to which the doctoral candidate is the main contributor. Its design was based on the aforementioned multilingual dataset. This corpus contains 424 texts classified into three levels of complexity and 11 domains. A fourth level will also be included, but the work is still in progress. Although this corpus is inspired by the multilingual dataset, some categories and subcategories may vary. This is due to the fact that in certain genres and subjects no texts written in Galician have been found. In addition, the corpus is smaller in size. The levels of difficulty of this corpus have been defined on the basis of the corresponding levels of difficulty established for Spanish, Portuguese and French. However, an adaptation was necessary to consider the specific linguistic aspects of Galician that affect the readability of the texts. This adaptation was done by taking into account the Celga[2] and CEFR[3] classifications for Galician. This corpus has not been published yet, but will be available soon.

[1] https://iread4skills.com/

[2] https://www.lingua.gal/o-galego/aprendelo/celga

[3] https://www.lingua.gal/c/document_library/get_file?folderId=1647060&name=DLFE-8921.pdf

7. Discussion

As research progresses, new questions arise about readability and complexity levels designed for text simplification or educational purposes, general or language-specific readability features, automated data transformation or validation of the classification. Some questions to be addressed in the future may include the following:

- Texts adapted for educational purposes and classified following the CEFR are commonly used to train and test automatic text classifiers based on complexity levels. Is there a correlation between readability levels and CEFR levels?
- Some linguistic phenomena, such as specific verb tenses or syntactic structures, affect the reading comprehension of a text. Focusing on the Galician case, which Galician-specific linguistic phenomena affect readability?
- In a readability scenario, what is the best data augmentation method for Galician?
- Considering that we are dealing with readability, is it possible to obtain high-quality resources by automatically transforming Spanish, Portuguese or French data into Galician?
- Regarding the classification of texts by readability levels, how can these classifications be validated? Is it possible to use generative systems to classify texts? If we use more than one annotator to validate the classification, how can we interpret the agreement between them?

Acknowledgments

This work has received financial support from the European Commission (Project: iRead4Skills, Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837), and from *Xunta de Galicia - Consellería de Cultura, Educación, Formación Profesional e Universidades* (Centro de investigación de Galicia accreditation 2024-2027 ED431G-2023/04), and the European Union (European Regional Development Fund - ERDF).

References

[1] D. Campos, P. Contreras, B. Riffo, M. Véliz, A. Reyes, Complejidad textual, lecturabilidad y rendimiento lector en una prueba de comprensión en escolares adolescentes, Universitas Psychologica 13 (2014) 1135–1146. URL: https://doi.org/10.11144/Javeriana.UPSY13-3.ctlr. doi:10.11144/Javeriana.UPSY13-3.ctlr.

[2] S. Vajjala, Trends, limitations and open challenges in automatic readability assessment research, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 5366–5377. URL: https://aclanthology.org/2022.lrec-1.574.

[3] F. Dell'Orletta, S. Montemagni, G. Venturi, Assessing document and sentence readability in less resourced languages and across textual genres, ITL - International Journal of Applied Linguistics 165 (2014) 163–193. doi:10.1075/itl.165.2.03del.

[4] R. Wilkens, D. Alfter, X. Wang, A. Pintard, A. Tack, K. P. Yancey, T. François, FABRA: French aggregator-based readability assessment toolkit, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 1217–1233. URL: https://aclanthology.org/2022.lrec-1.130.

[5] A. Quispesaravia, W. Perez, M. Sobrevilla Cabezudo, F. Alva-Manchego, Coh-Metrix-Esp: A complexity analysis tool for documents written in Spanish, in: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 4694–4698. URL: https://aclanthology.org/L16-1745.

[6] W. Xu, C. Callison-Burch, C. Napoles, Problems in current text simplification research: New data can help, Transactions of the Association for Computational Linguistics 3 (2015) 283–297. URL: https://aclanthology.org/Q15-1021. doi:10.1162/tacl_a_00139.

[7] G. Parodi, Corpus de aprendices de español (caes), Journal of Spanish Language Teaching 2 (2015) 194–200. doi:10.1080/23247797.2015.1084685.

[8] H. Saggion, E. Gómez-Martínez, E. Etayo, A. Anula, L. Bourg, Text simplification in simplext: Making texts more accessible, Procesamiento del Lenguaje Natural 47 (2011) 341–342. URL: https://www.researchgate.net/publication/277193726_Text_Simplification_in_Simplext_Making_Text_More_Accessible.

[9] S. Schwarm, M. Ostendorf, Reading level assessment using support vector machines and statistical language models, in: K. Knight, H. T. Ng, K. Oflazer (Eds.), Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 523–530. URL: https://aclanthology.org/P05-1065. doi:10.3115/1219840.1219905.

[10] S. Vajjala, D. Meurers, On improving the accuracy of readability classification using insights from second language acquisition, in: Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, NAACL HLT '12, Association for Computational Linguistics, USA, 2012, p. 163–173.

[11] A. Heintz, J. S. Choi, J. Batchelor, M. Karimi, A. Malatinszky, A large-scaled corpus for assessing text readability, Behavior Research Methods 55 (2022). doi:10.3758/s13428-022-01802-x.

[12] S. Vajjala, I. Lučić, OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification, in: J. Tetreault, J. Burstein, E. Kochmar, C. Leacock, H. Yannakoudakis (Eds.), Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 297–304. URL: https://aclanthology.org/W18-0535. doi:10.18653/v1/W18-0535.

[13] D. V. Ngo, Y. Parmentier, Towards sentence-level text readability assessment for French, in: S. Štajner, H. Saggio, M. Shardlow, F. Alva-Manchego (Eds.), Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023, pp. 78–84. URL: https://aclanthology.org/2023.tsar-1.8.

[14] G. Grego Bolli, D. Rini, S. Spina, Predicting readability of texts for italian l2 stu-

dents: A preliminary study, in: ALTE (2017). Learning and Assessment: Making the Connections – Proceedings of the ALTE 6th International Conference, 3-5 May 2017, Association of Language Testers in Europe, 2017, pp. 272–278. URL: https://www.researchgate.net/publication/320532498_Predicting_Readability_of_Texts_for_Italian_L2_Students_A_Preliminary_Study.

[15] E. Ribeiro, N. Mamede, J. Baptista, Automatic text readability assessment in European Portuguese, in: P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G. Oliveira, R. Amaro (Eds.), Proceedings of the 16th International Conference on Computational Processing of Portuguese, Association for Computational Lingustics, Santiago de Compostela, Galicia/Spain, 2024, pp. 97–107. URL: https://aclanthology.org/2024.propor-1.10.

[16] M. Martinc, S. Pollak, M. Robnik-Šikonja, Supervised and unsupervised neural approaches to text readability, Computational Linguistics 47 (2021) 141–179. URL: https://aclanthology.org/2021.cl-1.6. doi:10.1162/coli_a_00398.

[17] I. Gonzalez-Dios, M. J. Aranzabe, A. Díaz de Ilarraza, H. Salaberri, Simple or complex? assessing the readability of Basque texts, in: J. Tsujii, J. Hajic (Eds.), Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 334–344. URL: https://aclanthology.org/C14-1033.

[18] I. Madrazo Azpiazu, M. S. Pera, Is cross-lingual readability assessment possible?, Journal of the Association for Information Science and Technology 71 (2020) 644–656. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24293. doi:https://doi.org/10.1002/asi.24293.

[19] L. Vásquez-Rodríguez, P.-M. Cuenca-Jiménez, S. Morales-Esquivel, F. Alva-Manchego, A benchmark for neural readability assessment of texts in Spanish, in: S. Štajner, H. Saggion, D. Ferrés, M. Shardlow, K. C. Sheang, K. North, M. Zampieri, W. Xu (Eds.), Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Virtual), 2022, pp. 188–198. URL: https://aclanthology.org/2022.tsar-1.18. doi:10.18653/v1/2022.tsar-1.18.

[20] M. Bhargava, K. Vijayan, O. Anand, G. Raina, Exploration of transfer learning capability of multilingual models for text classification, in: Proceedings of the 2023 5th International Conference on Pattern Recognition and Intelligent Systems, PRIS '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 45–50. doi:10.1145/3609703.3609711.

[21] M. Rehan, M. S. I. Malik, M. M. Jamjoom, Fine-tuning transformer models using transfer learning for multilingual threatening text identification, IEEE Access 11 (2023). doi:10.1109/ACCESS.2023.3320062.

[22] A. Pintard, T. François, J. Nagant de Deuxchaisnes, S. Barbosa, M. L. Reis, M. Moutinho, R. Monteiro, R. Amaro, S. Correia, S. Rodríguez Rey, M. Garcia González, K. Mu, X. Blanco Escoda, iread4skills dataset 1: corpora by complexity level for fr, pt and sp, 2024. doi:10.5281/zenodo.10889888.