TOWARDS MULTIPURPOSE READABILITY ASSESSMENT

by

Ion Madrazo

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Computer Science

Boise State University

December 2016

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the thesis submitted by

Ion Madrazo

Thesis Title: Towards Multipurpose Readability Assessment

Date of Final Oral Examination: 27 October 2016

The following individuals read and discussed the thesis submitted by student Ion Madrazo, and they evaluated his presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Maria Soledad Pera, Ph.D.	Chair, Supervisory Committee
Timothy Andersen, Ph.D.	Member, Supervisory Committee
Nerea Lete, M.F.A.	Member, Supervisory Committee

The final reading approval of the thesis was granted by Maria Soledad Pera, Ph.D., Chair of the Supervisory Committee. The thesis was approved by the Graduate College.

Aitor, zuretzat.

ACKNOWLEDGMENTS

I would like to take advantage of this section to express my gratitude to all the people that in a more direct or indirect manner did this thesis possible:

I really appreciate the help received from all the **committee members**, and from the Computer Science department in **Boise State University**. You guided me to take the right path for succeeding in this Thesis.

Eskerrik asko Boiseko **Euskal komunitateko kide guztioi**, bereziki, Nere eta Izaskun, etxean bagina bezala sentituarazteagatik. Gure iritsiera eta egonaldia asko erraztu zen zuen laguntzari esker.

Eskerrik asko **Ixa taldeari**, bereziki Montse, Itziar eta Inigo, zuen tresnak erabiltzen utzi eta hauen inguruko duda guztiak argitzeagatik. Bertan egon nintzen jakintza oso baliozkoa izan zait tesi hau burutzerako orduan.

Thank you to all **lab mates**, specially to Nevena, Mikel and Iker, for making my daily life easier. I could not finish this thesis without the positive environments in our lab. Thank you for all the fun moments we had, and the hard moments where we supported each other.

Eskerrik asko **ama, aita eta Aitor** astero skypen bestaldin eoteatik. Hain urruti eotie goorra da danontzat, baine zuen babesa dauketela jakittiek asko laguntzeitt aurrea etten. Aitor tesi hau zuretzat da, holako bat ettea allako zeala seguru naolako. Eman eurre.

Sole, no hay palabras suficientes para darte las gracias por todo lo que has hecho por mi. Gracias por acogerme como estudiante y ser una profesora tan cercana,

estando siempre a nuestro lado en buenos y no tan buenos momentos. Gracias por toda la ayuda, han sido 4 semestres inolvidables, y espero que asi sean todos los que vengan. #SvenniesAlwaysMakeItTrue

ABSTRACT

Readability refers to the ease with which a reader can understand a text. Automatic readability assessment has been widely studied over the past 50 years. However, most of the studies focus on the development of tools that apply either to a single language, domain, or document type. This supposes duplicate efforts for both developers, who need to integrate multiple tools in their systems, and final users, who have to deal with incompatibilities among the readability scales of different tools. In this manuscript, we present MultiRead, a multipurpose readability assessment tool capable of predicting the reading difficulty of texts of varied type and length regardless of the language in which they are written. MultiRead bases its predictions on multiple indicators extracted from textual resources, including lexical, morphological, syntactical, semantic and social indicators. The latter are of particular interest given the recent adoption of social sites by users of different age and reading abilities. We gathered a leveled corpora comprised of textual resources in *English*, *Spanish* and *Basque* languages, with diverse length, source, domain and format. This corpora was used for assessing the effectiveness of MultiRead, and demonstrating that MultiRead outperforms other readability assessment systems, in terms of accuracy among all languages and document types evaluated.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

EN – English

EU – Basque

ES – Spanish

IR – Information Retrieval

NE – Named Entity

NLP – Natural Language Processing

PoS – Part of Speech

CHAPTER 1

INTRODUCTION

Reading is an important skill in the academic environment, a competence that can be critical for students' educational opportunities and their careers [94]. As reported by Lennon and Burdick [72], reading for learning takes place when the reader comprehends 75% of a text. This represents 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 [72]. Outside the educational environment, reading generally takes place for comprehension rather than for learning. In this context, it is critical to provide people with texts they can fully understand. For example, patients that properly understand documents disclosed to them before surgery are known to be less anxious before the operation and obtain more satisfactory results during posterior treatment [90]. Recent studies [63, 88, 90], however, 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. The reading level of a text is also influential from a marketing standpoint. The reports by Chebat et al. [36] demonstrate that the persuasiveness of advertisements is directly correlated to the readability of them. Furthermore, the authors in [36] argue that highly literate people would find low reading level argumentation naive, while low literate people would have trouble understanding complex arguments and

therefore lose interest in the ad. Along similar lines, Fang et al. [50], who analyzed the factors influencing the perceived value of attraction reviews in TripAdvisor, verify that the readability of a review is one of the most influencing factors for its perceived value. Whether for learning, understanding or ensuring persuasiveness, i.e., reaching a particular target audience, the complexity of texts to be read needs to be determined.

Every reader has different reading skills and the levels of difficulty of the texts they need depends also upon their personal objective. Therefore, providing institutions and readers with tools they can use to measure the complexity of a text so that they can assess whether it is adequate for a user is imperative. *Readability Assessment* tools[1] are certainly aimed for handling such a task by providing a mean to determine the degree of ease with which a reader can understand a given text, i.e. the *readability level* of the text.

Historically, teachers have been the main stakeholders of readability assessment formulas, using them to select new materials for their courses and curriculum design. However, lately, other stakeholders have found benefits in using readability assessment tools outside the academic environment. Automatic text simplification [95, 101], summarization for people with reading difficulties [51], book recommendation [89], literacy assessment [108], or legal and medical document complexity assessment [63, 85, 88, 90] are only a few examples of applications that take advantage of the complexity levels generated by existing readability assessment tools. Even in commercial environments, book publishers require professional linguistic services in order to tag their publications with a readability level required for their intended audience [12], a task that could similarly be completed by an automated tool.

[1]Readability assessment tool and readability assessment formula are used interchangeably in this document.

In estimating the complexity of texts, traditional formulas, such as Flesch [53], Dale-Chall [35], and Gunning FOG [17], became very popular in the late 1940's among educators for manually determining text difficulty. Most of these formulas relied on *shallow features*, which could easily be adapted to multiple languages and provide a simple way of determining text complexity. The multilingualism achieved by traditional formulas offered numerous benefits in contexts where the readability of more than one language was needed, i.e., book translation or learning a second language. However, traditional formulas were known to lack precision. For example, they could classify nonsense text as *simple to read*, just because it contained short and frequently-used words [43]. The insufficient precision encouraged researchers to study and develop better and more sophisticated methods for readability assessment that depended upon more in-depth text analysis [22, 56]. These new formulas continued taking advantage of shallow features, but incorporated more complex features based on the syntax and semantics of text. With the addition of new text complexity indicators, the tools became more precise, but at the same time more constrained regarding their language adaptability [29, 52]. In fact, the new tools used increasingly more language-dependent techniques, which made the systems unadaptable to estimate readability scores for texts in languages other than the one they were designed for. As a result, the multilingualism that was possible in early stages disappeared.

Versatility issues go beyond the number of languages that can be handled by readability assessment tools. They can also refer to the type of documents that can be processed by automatic readability assessment tools (see Figure 3.1 for sample documents that could be examined by readability assessment tools). While a wide variety of domains and document types have been studied in the literature, from single sentences [45] and books [46], to search results [40] and web-pages [109], most of the

existing tools focus exclusively on a specific domain, document type and language. This imposes additional efforts on stakeholders in the readability assessment area of study: developers, researchers, and final users (teachers, publishers, librarians, translators, and ad executives, to name a few). Typical issues that usually arise as result of having to use more than one readability assessment tool for the same task include (i) software integration costs, where developers need to spend more resources integrating multiple readability tools in their applications (ii) license costs, as companies need to pay for each tool license, (iii) prediction scale incompatibilities, where the predictions of each tool used do not match with each other, and (iv) longer learning periods, where final users need to learn how to use multiple tools instead of just one. Creating a multipurpose readability assessment tool would not only ease the creation of software that requires readability as a service, but would also facilitate the daily use of readability assessment tools for final users, who would no longer struggle using different assessment strategies.

Figure 1.1: Examples of document types for which MultiRead can determine readability.

With versatility and precision in mind, we designed and developed **MRAS**, a

Multipurpose **R**eadability **A**ssessment **S**ystem. MultiRead estimates the readability levels of all sorts of documents[2] in multiple languages, including text snippets, books, websites and even short and unstructured texts, such as the ones found in social media. MultiRead, which is open-source, is capable of detecting patterns that can influence the readability of a text in order to give a prediction of its difficulty, i.e., reading level. MultiRead adapts itself to the input text language and format and uses an adequate subset of features each time, creating, to the best of our knowledge, the first multipurpose readability assessment system.

For designing MultiRead, we explored features and methods used in the literature, designed novel features that positively influence the readability level estimating process, and analyzed how all those features can be adapted to be used in multipurpose readability assessment. In addition, given the increasing adoption of social media and social networking sites, which lead to the creation of new textual resources on a daily basis, from tweets and Facebook posts, to web-pages, we designed a novel set of features that take advantage of social elements such as hashtags, mentions or URLs. Finally, we analyzed strategies for combining all the aforementioned textual features, in order to predict the readability of a text.

During the research process, we also had to consider some technical aspects, such as the need of multilingual text processing tools and the lack of a dataset for development and evaluation purposes. Therefore, part of our research work involved exploring existing Natural Language Processing tools and identifying various textual resources that leaded us to the creation of a readability leveled corpora comprised of textual resources pertaining to multiple languages, domains and formats.

[2]In this manuscript the term *document* can refer to a text snippet, book, webpage, tweet, or any other textual representation for which readability will be predicted.

We conducted an in-depth study to validate the correctness of the design and effectiveness of MultiRead. We also assessed the benefits and limitations of a multipurpose readability assessment system. It is important to note that, for practical purposes, the proposed application has only been 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.

> “ *Write for the expert, but write so the non-expert can understand.* ”

Bernard Kilgore, *celebrated Wall Street Journal editor*

In writing this manuscript, we learned that, writing properly but in a manner the reader can understand is challenging but estimating the level of difficulty of a text is even more so. If not, consider this poem from G. A. Becquer, a famous Spanish post-romantic poem: “Volverán las oscuras golondrinas, en su balcón sus nidos a colgar...”(see Figure 1.2). A poem with simple words, yet full of meaning, cannot be assigned a level of difficulty comparable of that of a first grader (as Flesch-Kincaid does) or close to third grader (and Automatic Reading Index does). Only in its native Spanish, Flesch-Kinkaid assigns a grade level of 7 for the poem, which remains for third graders according to Automated Reading Index.

This is what we set out to accomplish in this research work, to be able to estimate a suitable level of difficulty, regardless of the language of the text, and regardless of its format: from long documents, to short tweets, to poems. How we did that, is detailed in the remainder of this manuscript: In Section 2, we discuss existing works

in the area of readability assessment, starting from the historically known traditional formulas to the current machine learning based readability assessment systems. We also discuss different strategies to measure readability assessment for texts in specific languages and document types. In Section 3, we describe the design of MultiRead, where apart for providing a background of the text processing techniques, we describe in detail each of the features specifically designed for MultiRead and the strategies used for combining those features. In Section 4, we present our evaluation framework as well of the results conducted to validate the performance of MultiRead and discuss the benefits and limitation of developing a multipurpose readability assessment tool.

In order to prove the validity of MultiRead in social related tasks we also developed an study to measure importance of the readability signal in the hashtag recommendation process, which we published in ACM RecSys 2016 [26]. We include this publication as an appendix (see Appendix A).

Figure 1.2: G.A. Becker's poem "Volveran las oscuras golondrinas" and its readability level estimated in original Spanish and translated English version.

Volverán las oscuras golondrinas	**The Dark Swallows Will Return**
Volverán las oscuras golondrinas	The dark swallows will return
en tu balcón sus nidos a colgar,	to your balcony to hang their nests
y, otra vez, con el ala a sus cristales	and, once again, with a wing to its glass
jugando llamarán;	playing, they'll call;
pero aquéllas que el vuelo refrenaban	but those that held back their flight
tu hermosura y mi dicha al contemplar,	when contemplating your beauty and my bliss,
aquéllas que aprendieron nuestros nombres...	those that learned our names...
ésas... ¡no volverán!	those... will not return!
Volverán las tupidas madreselvas	The dense honeysuckle will return
de tu jardín las tapias a escalar,	to your garden to climb its walls,
y otra vez a la tarde, aun más hermosas,	and, once again, in the evening, even more beautiful,
sus flores se abrirán;	its flowers will bloom;
pero aquéllas, cuajadas de rocío,	but those, studded with dew,
cuyas gotas mirábamos temblar	whose drops we watched tremble
y caer, como lágrimas del día...	and fall, like the tears of the day...
ésas... ¡no volverán!	those... will not return!
Volverán del amor en tus oídos	From love, to your ears will return
las palabras ardientes a sonar;	the ardent words to sound;
tu corazón, de su profundo sueño	your heart, from its deep sleep
tal vez despertará;	perhaps will awaken;
pero mudo y absorto y de rodillas,	but mute and engrossed and on its knees
como se adora a Dios ante su altar,	as God is adored before his altar,
como yo te he querido..., desengáñate:	as I have loved you..., undeceive yourself:
¡así no te querrán!	Like this, no one else will love you!

1 Flesch-Kincaid 7.7
2.4 Automated Readability Index 3.4

CHAPTER 2

BACKGROUND AND RELATED WORK

From the past six decades, different readability assessment systems have been developed [29, 52, 70]. In this section, we provide an in-depth discussion on readability assessment, from traditional formulas to state of the art techniques. We also discuss formulas applied to establish text complexity in different languages as well as for different types of texts.

2.1 Traditional Readability Assessment

Traditional readability formulas, such as Flesch [53], Dale-Chall [35], and Gunning FOG [17], make use of shallow features, mostly based on ratios of characters, terms, and sentences. Flesch [53] readability formula (see Equation 2.1) is based on a linear combination of the average length of words and average length of sentences in a document.

$$F = 206.835 - 1.015 \times (\frac{totalWords}{totalSentences}) - 84.6 \times (\frac{totalSyllabes}{totalWords}) \tag{2.1}$$

Kincaid et al. [69] adapted the Flesch formula to American education grade levels, in order to predict a grade level instead of a number between 0 and 100, as

the traditional Flesch formula does. This updated strategy, also known as Flesch-Kincaid (see Equation 2.2), uses the same features as Flesch, i.e., length of words and sentences, but combines them using different weights.

$$FK = 0.39 \times (\frac{totalWords}{totalSentences}) + 11.8 \times (\frac{totalSyllabes}{totalWords}) - 15.59 \tag{2.2}$$

Other alternatives, such as Dale-Chall [42] readability formula (see Equation 2.3), introduced the concept of simple and complex terms, taking advantage of a manually generated list[1] of 3000 easy terms. The frequency of those terms was used as an indicator of text complexity, together with the already known average sentence length.

$$DC = 15.79 \times (\frac{difficultWords}{totalSentences}) + 0.0496 \times (\frac{totalWords}{totalSentences}) \tag{2.3}$$

Similar to the Dale-Chall strategy, Gunning Fog [61] (See Equation 2.4) also considered the occurrences of simple or complex terms. However, instead of using a list to define complex terms, Gunning's readability formula considers a term complex if it has more than 3 syllables.

$$FOG = \frac{totalWords}{totalSentences} \times \frac{totalComplex}{totalWords} \tag{2.4}$$

SMOG [81] was developed as an improvement over Gunning Fog, in terms of precision. SMOG takes advantage of a non linear strategy (see Equation 2.5) that combines the total number of complex terms in a text and the total number of sentences. The method to determine whether a term is complex is the same as used by the Gunning Fog formula.

[1]The full list can be found in http://www.readabilityformulas.com/articles/dale-chall-readability-word-list.php

$$SMOG = 1.0430 \times \sqrt{\frac{30 \times totalComplex}{totalSentences}} + 3.1291 \tag{2.5}$$

Lasbarhet's index (see Equation 2.6), also known as LIX, predicts the difficulty to comprehend a text for a foreign reader. Similar to the previous formulas, it is based on the frequency of occurrence of complex terms per sentence. A term is considered as simple if it has less than 6 characters and the number of sentences is computed given the number of periods in the text.

$$LIX = \frac{totalWords}{totalPeriods} \times \frac{totalComplex}{totalPeriods} \tag{2.6}$$

The aforementioned formulas are a sample of the most popular among the hundreds available to date. Further details on existing traditional formulas (which are mostly based on sentence and term counts) can be found on the recent survey literature [29, 52].

The simplicity of these traditional formulas, made them easily adaptable to languages other than English. This is evidenced by the Spaulding's readability formula for Spanish [100], which uses the same two indicators as Dale-Chall's and Gunning Fog's readability formulas, i.e., ratio of difficult terms and average length of sentence in a document, with weights adapted to the Spanish language.

The aforementioned formulas are basic enough even to be computed manually, providing a simple way of estimating a text's complexity. A teacher or a librarian would compute the formula on the first pages of a book, and estimate whether the book was adequate for a reader without having to read the whole book. However, the formulas were shown to lack precision in some cases, such as the one described in [100], where completely nonsensical text was predicted to be easily readable by the

aforementioned readability formulas. As an example, the phrase *sv eni sar ein de er*, would be considered as easily readable by all the aforementioned readability formulas, just because it has short terms, even if it is completely nonsensical, or the term *quark* which is considered simple by most of the traditional readability formulas, due to its length, despite being a high level technical term [107].

The increase in popularity of machine learning techniques and the need to improve predictive quality of traditional formulas lead the readability assessment into a new era of study. An era where readability formulas take advantage of supervised learning techniques to combine tens or hundreds of indicators. Even if shallow features are still included in current readability assessment tools, they are usually considered baseline features, and features that consider other language aspects, such as syntax or semantics, play a more significant role [29].

2.2 Readability Assessment by Languages

Adapting readability assessment tools to several languages have been the main focus for researchers on recent years. This is evidenced by the fact that there exist at least one prediction formula for each of the most popular languages spoken worldwide. A description of representative and recent tools designed and developed to predict difficulty levels for texts in different languages is presented below.

For **English**, the readability assessment system presented in [22] predicted only two levels of difficulty, simple or complex, using elaborate features, such as ambiguity among terms in the texts. Other authors [51], 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. The readability

prediction system for financial documents presented in [32] was based on features such as the presence of active voice or number of hidden verbs. It is also important to mention two commercial readability assessment tools, Lexile[2] and AR[3], which are widely-used among academic professionals in the USA, being used by over 150 publishers [12] . Even if their algorithms are not public, they are known to use shallow features showing how common terms of a text are and how long sentences are [72]. The literature pertaining to readability for text in English is abundant. For a more in-depth discussion on readability assessment for texts in English refer to [29, 52].

In contrast to English, **Spanish** readability assessment has not seen any significant improvement regarding features in recent years, as most of the existing works are still based on shallow features. Among the well-known readability assessment tools for Spanish, SSR [100] is based on the analysis of sentence length and number of infrequent words per sentences, whereas LC and SCI [24] were based on density of low frequency words in text. Other systems [47, 102] incorporate strategies to combine the aforementioned methods to improve readability estimation.

Compared to other languages, **Basque** readability assessment is reduced to only one system. Due to the fact that Basque is considered a minority language and shares little similarity with most spoken languages, limited research has been done in the area. So far, ErreXail [60] is the only system created for Basque readability assessment. ErreXail was developed to predict two different readability values, simple or complex, using features mostly based on ratios of common natural language processing labels, such as Part-of-Speech tags or morphology annotations.

Similar to Basque, the literature for **Arabic** readability assessment is also very

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

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

recent. Al-Ajlan et al. [16] developed a readability assessment tool based on only two features: average letters per term and average terms per sentence. These features were analyzed using a Support Vector Machine in order to classify text as simple or complex. Recently, J. Forsyth [55] performed another study with a significantly larger amount of features than previous studies, demonstrating the validity of lexical and discourse features for Arabic readability assessment. In a simpler approach, Mahmoud et al. [48] presented a modification of the Flesch readability. Apart from the common Flesch indicators, this formula also includes information about short, long and stress syllables, as well as some other textual aspects that are only found on formal texts.

For **Italian** and **Russian**, the research conducted by F. Dell'Orletta [45] and Karpov et al. [67], respectively, demonstrated the importance of structural features for readability prediction. Both research works are focused on a combination of several syntactic features, including features that measured the complexity of syntactic trees.

Unlike readability assessment tools for English, Spanish, and Italian, to name a few, structural features do not seem to have such a positive influence for **Chinese**. Therefore, most of the research works related to Chinese readability assessment have been focused only on lexical features, such as Tf-Idf of terms [37, 41].

Rather than focusing on the general reader, François and Fairon [56] developed a system for **French** with 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. Uitdenbogerd [106] also focused on the same task. However, his study only focused on English natives that learned French. As a novelty, he proposed a feature that considered the occurrence of true cognates, terms that were same or similar in both languages, since those terms are

the ones than this audience easier learns. Wang [107] also proposed the use of true cognates for readability assessment, developing an automatic true cognate identifier.

2.3 Readability Assessment by Document Type

Traditional readability assessment has usually been oriented to relatively long text snippets [17, 35, 53]. While state-of-the-art [29, 55, 60] is also mostly oriented to such type of text, recent studies also explore methods for assessing the readability of other types of document.

Several studies are focused on the analysis of readability for single sentences [45, 67]. Most these studies are usually part of text simplification systems, which use readability assessment for choosing which sentences need simplification. Dell'Orletta et al. [45] developed a readability assessment tool for Italian sentences, combining lexical, syntactical and semantic features. Karpov et al. [67] developed a similar study for Russian, making a big emphasis on syntactical features. Both studies concluded that structural features have the most relevance for sentence readability prediction.

Web-related readability assessment has also been studied. Web pages are usually challenging for readability assessment given their varied topics and formats. Collins-Thompson et al. [40] assessed the readability of search results by considering information in both the title and the snippet retrieved by the search engine, and the full content of the pointed web page. The authors take advantage of language models for predicting readability, since these models are the most adequate for predicting the readability of short and noisy texts, such as web pages and their snippets [40].

Yu et al. [109] presented a Firefox plugin to automatically enhance the readability of web page for Asians that do not speak fluent English. For doing so, their systems

considered several structures known to be complex for non-English native speakers and applied several transformation to make them more readable.

Along similar lines, Kanungo et al. [66] developed a readability assessment tool for search result summaries. Their system combines several traditional readability formulas, such as Flesch [53] or Gunning-FOG [17], with some novel features specifically designed for their task. The latter refer to features that measure the number of strange characters or repeated keywords, in order to detect spam summaries. This is an important aspect for these type of documents, because spammers try to trick search engines with summaries full of keywords, that are usually recognized as simple by readability assessment tools [66].

Even if books also contain long snippets, which traditional readability formulas are able to handle, copyright regulations make books contents difficult to obtain. In order to overcome this issue, Dening et al. [46] presented a readability assessment tool for K-12 books, which goes beyond traditional textual contents. Their system focuses on available book metadata such as its author or genre.

2.4 Applications of Readability Assessment

Educational applications have traditionally been the main focus for readability formulas. Popular tools such as Lexile and AR were specifically designed to help teachers and librarians select books for children. Both systems are currently commonly used by book publishers to catalog books given their readability level [12].

Readability assessment tools have also been included in automatic book recommendation systems. Rabbit [89] makes book recommendations oriented to K-12 children considering multiple criteria that include several appealing factors for the

reader as well as the readability score for the recommended books. This permits Rabbit to not only recommend books that are of interest of to a reader, but also ensure that he is going to be able to understand them.

Text simplification also takes advantage of readability assessment tools [45, 51]. Knowing if a text needs to be simplified is an important prerequisite for such a system [45]. More specifically, being able to recognize which parts of a text are the ones making the text complex is also important. Single sentence readability assessment [45] has been used to handle both issues. Text simplification can be seen as an iterative process where a text can be infinitely simplified. For this task, knowing when to stop is also a must. Therefore, readability assessment has also been used to determine if a simplification was enough or not, both as an evaluation method or stopping criteria [102].

In some contexts, such as the medical domain [90] or food diseases [49], it is critical to provide people with documents that they can fully understand. For example, patients that properly understand documents disclosed to them before surgery, are known to be less anxious before the operation and obtain more satisfactory results during posterior treatment [90]. Therefore, some institutions are currently enforced by law to ensure that the documents they generate match the reading level of average people [30, 57]. Several studies [30, 57, 63, 88, 90] have been conducted to assess whether this enforcement is fulfilled. However, most of them show that documents that are supposed to be suited for average readers, tend to be too specialized and even well-educated adults have trouble understanding them [30, 57, 63, 88, 90].

Web search engines are increasingly getting more personalized towards their users. With the goal of providing users with resources that are both of interest and match their level of understanding, several applications have incorporated a readability

signal in their systems. Collins-Thompson et al. [41] presented a re-ranking strategy based on the readability of retrieved documents, so that the ones that were more adequate to the user would be ranked higher. On the other side, Kanungo et al. [66] took advantage of readability assessment for improving the way summary snippets where created in Yahoo!. This method was oriented to make the summaries more readable so that users would better know what to expect when they clicked on each result.

Even if the application domains discussed in this section are the most prominent, they are not he only ones that benefit from high-quality readability assessment. Other applications such as, translation [62] or dyslexia-related studies [98] also take advantage of such complexity assessment.

2.5 Feature Fusioning Techniques in Readability Assessment

When estimating the readability level of a document, analyzing features in isolation is not enough. Instead, it is important to generate a single score that simultaneously considers the information captured in each individual feature. This leads to a more well-rounded assessment of the document and thus a better estimation of its level of difficulty.

In addressing the feature fusioning challenge for readability assessment, many diverse techniques have been considered [29]. Collins et al. [41] used a naive Bayes model, while Denning et al. [46] took advantage of a linear regression for combining features. Francois et al. [56] determined the logistic regression to fit best with their system. However, most of state-of-the-art systems have used Support Vector Machines [52, 105, 96, 29] making this technique the most popular in the area.

2.6 Readability Assessment and Multilingual Needs

While the number of readability assessment systems oriented to individual languages is high, little research has been done regarding multilingualism. To the best of our knowledge, the study carried by De Clercq et al .[44], is the only work that handles more than one language for readability prediction. In this work, readability assessment techniques are analyzed for both Dutch and English, and a readability level prediction tool is developed for each language. The research work focuses on comparing which features are valuable for each on the languages analyzed, concluding that the best feature set for both languages is significantly similar. This fact supports the development of MultiRead, since it proved that a unique set of features could be used for automatically predicting the readability in several languages.

MultiRead is distinct from De Clercq et al.'s work in the fact that MultiRead is only one system, with a comprehensive set of features to be able to predict the readability of multiple languages, instead of building one system for each language as they do. Furthermore, De Clercq et al. only focus on the readability prediction of long text passages, while the set of documents that MultiRead can handle is broader: MultiRead can process documents that vary in length (long and short snippets), format (web pages, tweets, or plain text), and topic.

CHAPTER 3

METHOD

MultiRead is an assessment tool capable of automatically predicting the readability level of any given document,[1] regardless of its type, format, length or language. To do so, MultiRead takes advantage of several tools (described in Section 3.1), as well as various text processing strategies (described in Section 3.2). An overview of its design methodology is described in Sections 3.3 - 3.6.

3.1 Tools

Whether for processing documents or extracting features from a document that can be analyzed in order to predict its readability level, MultiRead depends upon several existing tools and techniques, which we describe below.

3.1.1 NLP Toolkits

In designing and developing MultiRead we analyzed a number of existing NLP toolkits. The goal of this analysis was to find one than fulfilled the text processing requirements of MultiRead, regarding both the number of languages and functionalities handled by our tool.

[1] As previously state, in this manuscript the term document can refer to a snippet, book, webpage, or tweet, to name a few.

We focused our assessment on the toolkits presented in Table 3.1, since they are popular, well-documented, and continuously updated.

Table 3.1: Comparison of existing NLP tools in terms of languages they can handle and basic text processing functionalities they provide.

Tool	Languages	Tokenization	Stemming	PoS	Chunking	Dependency	Named Entity
NLTK	English +	Yes	No	Yes	Yes	In development	Poor
CoreNLP	English +	Yes	Some languages	Yes	Yes	Some languages	No
OpenNLP	English +	Yes	Yes	Yes	Yes	Yes	Yes
SyntaxNet	70+	Yes	No	Yes	No	Yes	No
Freeling	14	Yes	Yes	Yes	Yes	Yes	Yes
Katea	Basque	Yes	Yes	Yes	Yes	Yes	Yes
TwitterNLP	English	Yes	No	Yes	No	Yes	No

NLTK [31], OpenNLP [1] and CoreNLP [80] were originally designed for English, with the possibility of extension to other languages. Unfortunately, these toolkits currently offer support for a limited number of languages and expanding them to new languages would require training models specifically tailored to each of them, which is a non-trivial task.

Instead of using specific language models, SyntaxNet [2, 23] is compatible with already existing commonly-used model formats, such as the one used in the Universal Dependency Treebank [3], which contains models for more than 70 languages. However, given the recentness of its development, SyntaxNet still does not offer support for some common NLP tasks, i.e., tokenization or lemmatization, that are imperative for the readability prediction process.

Freeling [86, 87] is a project more consolidated than SyntaxNet, offering a good balance between languages supported and functionalities. In addition to tokenization functionalities, Freeling offers Part-of-Speech tagging, syntactic parsing, dependency parsing and semantic labeling capabilities, and supports 14 languages: Asturian, Catalan, German, English, French, Galician, Croatian, Italian, Norwegian, Portuguese, Russian, Slovene, Spanish, and Welsh.

Katea [14, 20, 28] is the only existing set of NLP tools developed for Basque. Katea is composed by Morpheus [14] (morpho-syntactic analysis), eustagger [20] (lemmatization and syntactic function identification), eihera [19] (named entity detection), ixati [14] (shallow parsing) and maltixa [28] (dependency parsing).

We also considered TwitterNLP [58], a natural language processing toolkit specifically designed for analyzing Twitter-generated data, i.e., tweets. This toolkit provides support for NLP tasks such as tokenization or dependency parsing. However, it only works on tweets written in English and not all functionalities needed by MultiRead are provided. Consequently, we decided not to include it as part of our tool.

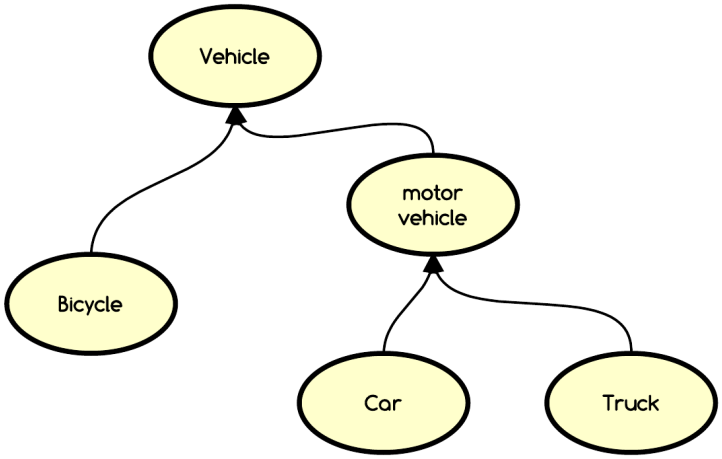
Offering support for the three languages considered in this thesis (English, Spanish and Basque), SyntaxNet would be the ideal choice of tool for text processing. However, given that it still does not offer some basic functionalities, such as tokenization or shallow parsing, we decided to use two other tools that complement each other: Freeling and Katea. Together they fullfill both the language and functionality needs of MultiRead.

3.1.2 WordNet

Wordnet [83] is a lexical database for English where terms, i.e., nouns, verbs and adjectives, are grouped into sets of synonyms, also known as *synsets*. Synsets are related to each other by several semantic relationships, such as hyperonymie or hyponimie, forming a tree structure. This structure eases some basic operations, such as measuring distance or similarity between synsets.

As an example of WordNet's tree structure, Figure 3.1 illustrates a partial neighborhood for the *motor vehicle* synset. This synset has *vehicle* as hypernym and *car* and *truck* as hyponyms.

Figure 3.1: Example of a reduced tree generated around the *vehicle* concept in Wordnet.



For MultiRead we used the WordNet implementation developed by the IXA NLP group [15], which provides support for five languages including English, Spanish and Basque. This implementation takes advantage of the Inter-Lingual Index, which offers a simple way to map synsets in different languages that have the same meaning to one unique identification code.

3.2 Text Processing

Natural language text is generally unstructured, making it hard to handle for computers [78]. Therefore, the aim of text processing is to infer inherent structures in text in order to ease consequent processes. In the remaining of this section, we discuss basic text processing operations applied by MultiRead.

3.2.1 Tokenization

Tokenization is the process of splitting a text into smaller parts, i.e., tokens [79]. A token represents each sensical part of a text, which usually corresponds to a term, a number, or a punctuation mark.

Even if tokenization may look as simple as dividing text by spaces and punctuation marks, usually it is not as trivial, as each language has specific punctuation rules that need to be considered [104].

As illustrated in Table 3.2, tokenization goes further than splitting a sentence based on spaces. Both the words *aren't* and *student's* need to be separated into two tokens in English, while *1991* and the *period* need to be kept together to show ordinality in Basque numbers. The reason for this different tokenization is that *are* and *not* are two separate terms with full meaning on their own in English, whereas *1991* would lose its ordinal sense if the period is removed in Basque.

A precise tokenization is important, since most of the processes involved in identifying features from text are usually preceded by this analysis.

Table 3.2: Tokenization of sentences in different languages.

Basque	**Sentence**	1991. urtean jaio zen.						
	Tokens	1991.	urtean	jaio	zen	.		
Spanish	**Sentence**	La O.T.A.N. ha firmado el acuerdo.						
	Tokens	La	O.T.A.N.	ha	firmado	el	acuerdo	.
English	**Sentence**	Those aren't student's tables.						
	Tokens	Those	are	n't	student	's	tables	.

3.2.2 Stopword removal

Stopword removal or stopping refers to the process of removing stopwords from a text [79]. A stopword is a term that does not add any important information to

the task that is performed, usually creating unnecessary noise that hinders valuable information in a document. While common stopwords include prepositions and articles, the frequency of occurrence of a term is also a good indicator for stopwords since, in general, the more frequently a term appears in a text o corpus, the less information it provides about the text. Some examples of general purpose stopwords are *a*, *the* or *is*. However, depending on the domain, terms such as *computer* or *algorithm* can be treated as stopwords among computer science documents, given their high frequency of occurrence among documents in this domain.

The purpose of stopword removal is usually two-fold: speeding up later processes and reducing noise in text. This process is usually performed without the need of any specific tool, just using a stopword list. In MultiRead, we take advantage of a popular stopword list for 50 different languages [9].

3.2.3 Stemming/Lemmatization

The goal of both stemming and lemmatization is to achieve a normalized version of a term [79]. Stemming and lemmatization differ in the way the normalized form is obtained. While lemmatization is able to achieve real canonical form (i.e., lemma) of a word (the one appearing in the dictionary), stemming simply chops common prefixes and suffixes to obtain an approximation of the lemma.

Stemming or lemmatization is usually useful for search and comparison tasks as it reduces the search space among all terms. As an example, in calculating the frequency of occurrence of the verb *play*, it is more representative to count all word-forms (*play*, *plays*, *played*) at the same time, than to separate each word-form frequency.

When both techniques are available, MultiRead favors lemmatization over stemming, since the former yields a smaller term space. Unfortunately, lemmatization

techniques are not available for some languages, given the complexity of this process. Freeling is used for lemmatization in Spanish and English, while Katea serves the same purpose for Basque. Several examples of lemmatization can be seen in Table 3.3.

Table 3.3: Example of lemmatization for different languages.

	English			**Spanish**			**Basque**		
Original	plays	are	won	fuiste	vine	comieron	nator	balituzte	dakit
Lemmatized	play	be	win	ir	venir	comer	etorri	edun	jakin

3.2.4 Part of Speech Tagging

Part of Speech (PoS) tagging is the process of labeling each token with a tag that represents the function the token has in a sentence [79]. PoS tags usually differ from language to language[2], however, the most predominant tags, such as verb, adjective or noun, are common among most languages. As shown in Table 3.4, for the sentence "Did they win the race?" most of the PoS tags used are similar, regardless of the language in which the sentence is written, with the exception of some auxiliary verbs and participles.

PoS tags are the building blocks of structure in text. They can be used alone for applications that need to consider very basic structure in text or need to better identify terms, since a term can have different meaning depending on its PoS. For example, *left* means *leaving* if it is treated as a verb, but it means the opposite of *right* when it treated as a noun.

[2]The Penn Treebank project defines 36 PoS tags for English, which can be seen here https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html whereas the Ancora project defines 49 PoS tags for english which can be seen here: http://clic.ub.edu/corpus/webfm_send/18

Table 3.4: Part of Speech tagging for the sentence “Did they win the race?” in multiple languages.

English	Did	they	win	the	race	?
	Verb (Auxiliar)	Pronoun	Verb	Determinant	Noun	Symbol
Spanish	Ganaron	(ellos)	la	carrera		?
	Verb	Pronoun (Elliptic)	Determinant	Noun		Symbol
Basque	Lasterketa	irabazi	al	zuten	(haiek)	?
	Noun	Verb	Particle	Verb (Auxiliar)	Pronoun (Elliptic)	Symbol

3.2.5 Named Entity Detection

A named entity is a token or group of tokens that represent a known entity, such as a person, a location, or an organization [79]. Depending on the complexity of the tool that performs this analysis, those entities can also be linked to a knowledge base, such as DBpedia [71] where more structured information about the entity can be found.

A named entity is usually a relevant term in a text, something that a user may be interested in or looking for, i.e., a name of a location or person. For this reason, they are often treated as key points, i.e., terms that are representative of the content of a document, for searching or indexing purposes. For an example of named entities identified using Freeling, see Table 3.5.

Table 3.5: Named entity detection of the sentence “Salvador Dali was born in Figueres”

Sentence	Salvador	Dali	was	born	in	Figueres	.
Named Entity	person	person				location	

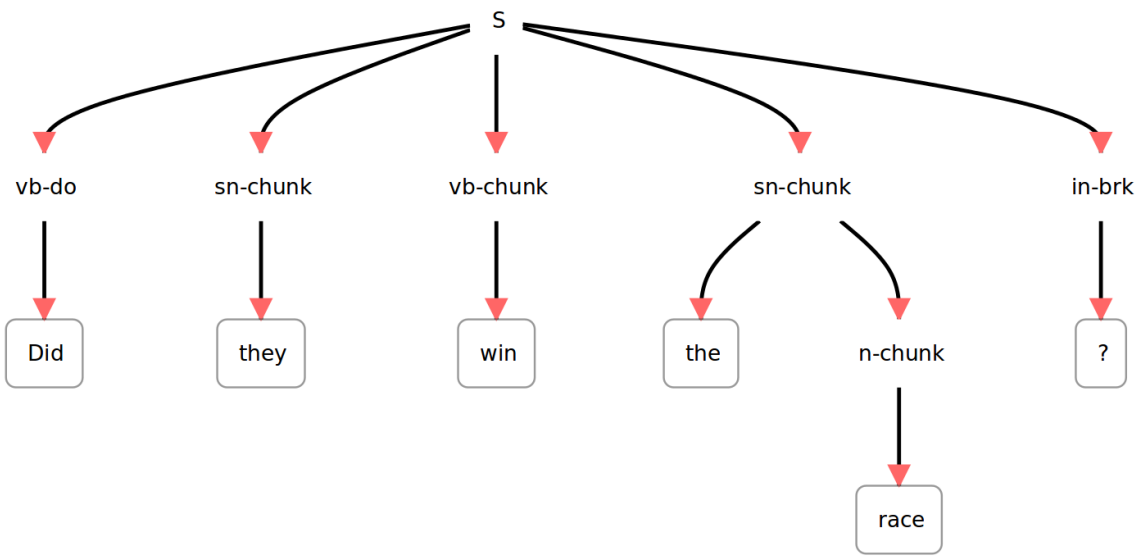
3.2.6 Shallow Parsing

Shallow parsing, also called chunking, refers to the process of grouping tokens into chunks [79]. A chunk usually consists of a small phrase of about 1 to 4 terms. The terms in each chunk are somehow connected to each other and together express a

senseful concept. There are two types of chunks, depending on if they express a noun (*sn-chunk*) or a verb phrase (*vb-chunk*). An example of a shallow parsing of a sentence can be seen in Figure 3.2. Note that we only provide an example in English, since this shallow parsing works in the same manner for most languages.

Shallow parsing is useful for tasks that require using bigger chunks of text than just tokens. One simple example could be autocompletion, where full senseful chunks can be suggested to the user instead of just separate tokens. Named entity recognition also benefits from shallow parsing, since most of the named entities are usually exact noun-chunks.

Figure 3.2: Shallow parsing of the sentence "Did they win the race?" where verb chunks are denoted with a *vb* prefix and noun chunks with a *sn* prefix.

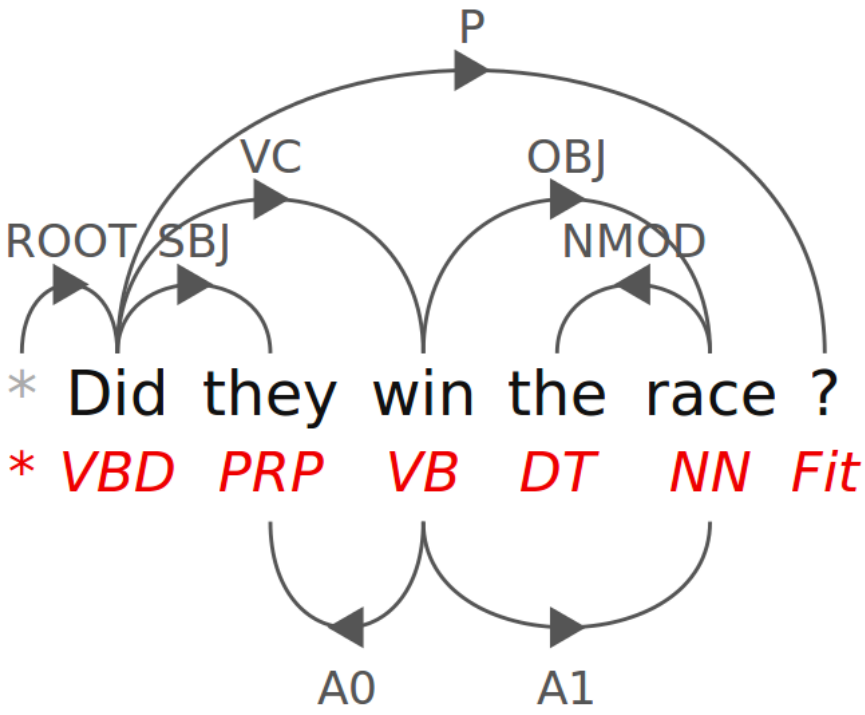


3.2.7 Dependency Parsing

Dependency parsing goes further than shallow parsing by establishing relationships between tokens rather than just grouping them [79]. Given these relationships, a

dependency tree is generated. This tree usually has a root node representing the main verb of the sentence, which has the subject and objects of the sentence as children. An example of a dependency parsed sentence is shown in Figure 3.3. Once again, only one language is presented in the example, since dependency parsing is very similar for the languages considered in this research work.

Figure 3.3: Dependency parsing of the sentence "Did they win the race?", where the root node is the term *did* with the term *they* as subject and the term *win* as verb.



One of the benefits of dependency parsing over shallow parsing is the possibility to choose a granularity when selecting chunks with full meaning. Using the tree generated by the dependencies, any cut will cause the sub-trees to be fully sensical. For example, cutting the tree depicted in Figure 3.3 in the node *did*, would create chunks with full sense *they* and *win the race* while cutting the tree in the node *win*

would create *the race*. Because of this possibility, dependency parsing is highly used in language transformation applications, such as transforming a statement into a question.

3.3 Design Overview

MultiRead is based on a supervised learning approach that relies on knowledge acquired from a set of text given their readability. MultiRead is capable of identifying patterns in texts in order to predict the readability of any new document D. In designing and developing MultiRead, we followed the steps illustrated in Figure 3.4 and discussed in the following sections.

3.4 Adapting to the Input Document

One of the main features of MultiRead is its versatility, since MultiRead is capable of predicting readability levels for documents of different format, length and language. Consequently, not all documents handled by MultiRead can be treated in the same manner.

The language in which a document is written is the characteristic that most influences its processing, as it determines the text processing toolkit that will be used for it. For doing so, MultiRead takes advantage of Freeling's language identification module. While this module is limited to the languages supported by Freeling, if MultiRead were to be expanded to be applicable to other languages, developing a language identification module would be relatively trivial, as a simple bigram distribution would provide enough accuracy for this task.

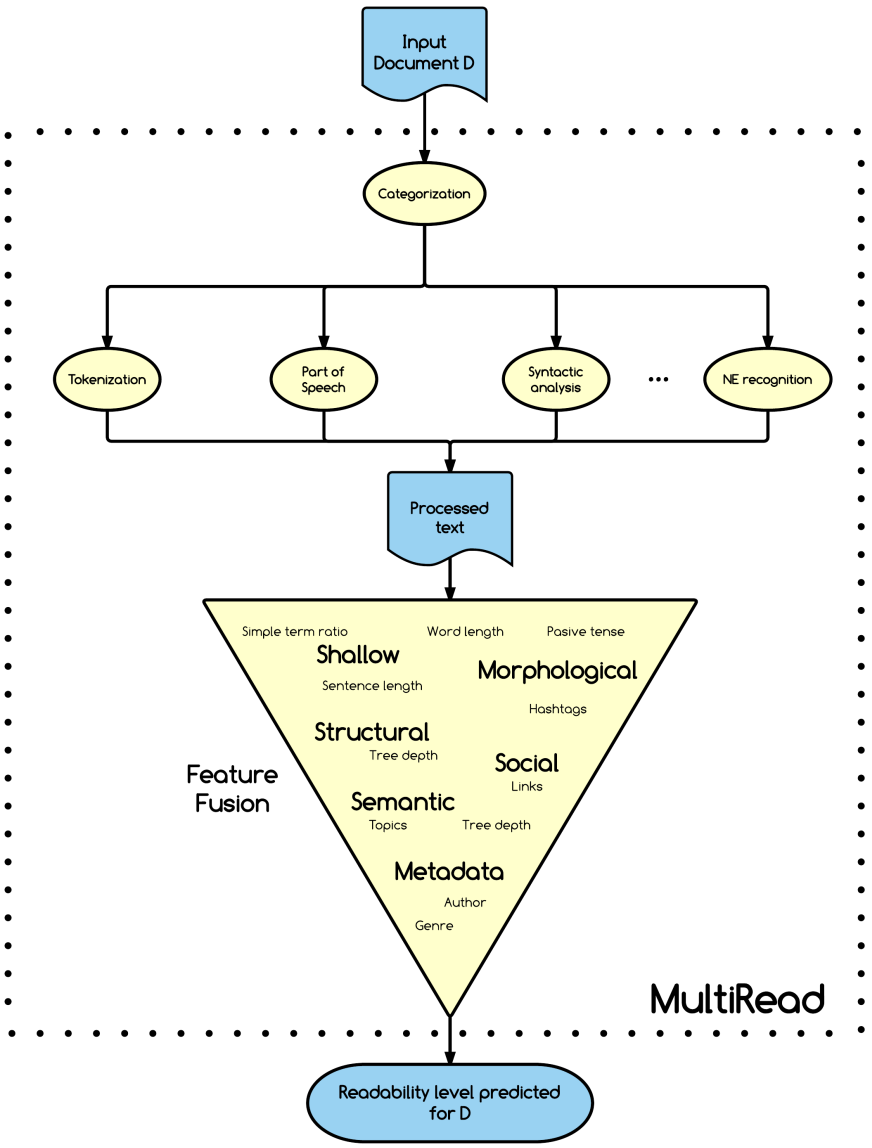


Figure 3.4: Overview of the design of MultiRead

As opposed to language, the format or length of a document are not as discriminating as the language, in terms of internal processing steps. In other words, MultiRead applies the same processing steps for any document D, regardless of D's type, format, or length, in order to extract and infer as much information as possible about D for

better estimating its degree of difficulty.

3.5 Feature Extraction

A feature is a numeric representation intended to capture information about a text from different perspectives. Feature engineering, i.e., identifying patterns that influence the readability of a text, is one of the most important aspects of this thesis. A good feature set determines the quality of a classifier, and therefore the quality of a readability assessment system. In the rest of this section, we provide a description of each feature examined in MultiRead.

3.5.1 Shallow Features

Shallow features [17, 35, 53] have historically shown to be of good use when predicting readability. Even if they sometimes lack precision [43], they serve as a good baseline for readability assessment systems. We describe below shallow features considered by MultiRead.

Word length

Everyday terms are usually short in most languages, they are preferred for spoken language over their longer synonyms, in order to maintain more fluent conversations. On the other hand, difficult terms are longer the more technical or scientific they are. Therefore, short terms are the ones youngsters first learn and better comprehend. This leads to hypothesize that term length can be correlated with the readability of a document. To take advantage of this, MultiRead depends upon four features

(computed as in Equation 3.1) that focus on term length: *average length of terms* and *average length of their lemmas* in D, with and without stopwords for both cases.

$$WordLength(D) = \sum_{w \in D_w} \frac{length(w)}{|D_w|} \tag{3.1}$$

where w is a word in D, $|D_w|$ is the total number of words in D and $length(w)$ is a function that calculates the number of characters in w.

Sentence Length

The length of a sentence can have direct correlation with its difficulty. Sentences oriented to people with low understanding skills are usually short as they just focus on simple ideas and facts with very naive argumentation. On the other hand, documents oriented to more technical or complex audiences contain more argumentation. Consequently, more sub-clauses are incorporated in the sentences resulting in longer sentences. MultiRead takes advantage of this fact analyzing two features, *average sentence length in D* with and without stop words, each of which is computed using Equation 3.2.

$$SentenceLength(D) = \sum_{s \in D_s} \frac{length(s)}{|D_s|} \tag{3.2}$$

where s is a sentence in D, $|D_s|$ is the total number of sentences in D and $length(s)$ is a function that calculates the number of word in s.

Ratio of Simple Terms

The vocabulary of a text strongly determines its readability [18, 44, 91]. Everyday terms tend to be easy for readers to understand, while more technical, domain specific,

or abstract terms tend to be more difficult. The frequency of occurrences of simple and complex terms have been demonstrated to be positively correlated with readability levels [42]. Consequently, using Equation 3.3, MultiRead computes a score that makes it possible to examine the *ratio of occurrence of simple terms* among all the terms in D. Determining whether a term is simple or complex, however, is still an open task. Therefore, MultiRead adopts two different , yet well-established, strategies to determine if a term is simple:

- **Lookup Table** Dale-Chall [42] created a list of 3000 terms considered simple. This list is used as one of the techniques to detect whether a term is simple or not.

- **Length** Gunning [61] simply considered terms that contained more than 3 syllables as complex. This technique is also incorporated in MultiRead.

$$RatioOfSimpleTerms(D, t_{simple}) = \frac{|Simple(D, t_{simple})|}{|D_w|} \tag{3.3}$$

where $|D_w|$ is the total number of words in D and $|Simple(w, t_{simple})|$ is the total number of simple words in D, as determined by a given strategy, i.e., t_{simple}, which refers to either Dale-Chall or Gunning Fox methodologies.

Ratio of Different Terms

As mentioned before, vocabulary is an aspect that strongly correlates with the readability of a text. As a hypothesis, a document that has very diverse vocabulary might denote an effort of the author to make a text more sophisticated, since experienced writers usually try to avoid repetitions using synonyms and different rephrasing. We

take advantage of this phenomena by measuring the *ratio of terms that only occur once* with respect to the ones that occur more than once in D.

$$RatioOfDifferentTerms(D) = \frac{|Unique(D_w)|}{|D_w| - |Unique(D_w)|} \tag{3.4}$$

where $|D_w|$ is the total number of words in D and $|Unique(D_w)|$ refers to the total number of words in D that only occur once, after lemmatization.

3.5.2 Morphological Features

Morphological features capture how terms are formed from their root. Even if this aspect is not relevant for some languages, including English, it has been shown to be a strong predictor for readability scores in morphology-rich languages such as Basque [60]. Morphological features analyzed in MultiRead are described as follows.

Inflection Ratio

Inflection is the modification of a term to express different grammatical categories, including case, tense, person, number, gender, mood or voice. For example, the verb *play* can be inflected to *played* to indicate past tense. Even if verbs in English have very few inflection forms (e.g. speak, speaks, spoke, spoken) some languages, such as Spanish or Basque, use inflection very extensively, taking advantage of most of the aforementioned categories. As an example, the Basque auxiliary verb **edun*[3] (one of the verbs with highest occurrence in Basque) has over 1000 inflection forms[4]. The longer the inflected form, the more complex it tends to be and the harder to learn and

[3]The star in front of **edun* is a notation used in verbs that its canonical form was arbitrarily created and therefore never occurs in a text.

[4]See simplified *edun verb building table at
https://nereagaite.files.wordpress.com/2011/11/nor_nori_nork_full_table.png .

understand for people with low proficiency of the language. To take advantage of this fact, MultiRead considers to what extent are terms inflected in D and measures the *average ratio between the character length of word-forms and their respective lemmas.*

$$InflectionRatio(D) = \frac{\sum_{w \in D} \frac{length(lemma(w))}{length(w)}}{|D_S|} \tag{3.5}$$

where w is a word in D, $|D_w|$ is the total number of words in D, $lemma(w)$ is a function that yields the lemma of w, and $length(lemma(w))$ is the character length of the lemma of w.

Morphological Phenomena Frequencies

Some morphology phenomena are more frequent than others in everyday language, whereas some phenomena only happen on high level structures of text. As an example, subjunctive mood (*if I were*) is less frequently used than indicative (*I was*) tense, and is more common in higher level texts. The same happens with the Nor-Nori (Who-ToWho) person form in Basque, a form that is usually avoided in spoken language, and is only used by proficient writers. To take advantage of the described phenomena and to discover other similar phenomena, MultiRead examines the frequency of occurrence of each morphological phenomena t_{morph} in D. Using Equation 3.6, MultiRead estimates the *ratio of occurrence of each phenomena* with respect to the number of tokens in D. Phenomena analyzed by MultiRead include:

- **Case** (English:0,Spanish:0,Basque:17) Ablative, Absolutive, Adlative1, Adlative2, Adlative3, Dative, Destinative, Descriptive, Ergative, Genitive1, Genitive2, Inesive, Instrumental, Motivativ, Partitiv, Prolativ and Sociativ.

- **Tense** (English:3,Spanish:3,Basque:3) Past, Present, Future.

- **Person** (English:0,Spanish:0,Basque:4) Who, Who-Who, Who-ToWho, Who-ToWho-What.
- **Number** (English:2,Spanish:2,Basque:2) Singular, Plural.
- **Gender** (English:0,Spanish:2,Basque:0) Masculine, Feminine, Neutral.
- **Mood** (English:3,Spanish:3,Basque:4) Indicative, Subjunctive, Imperative, Hypothetical.
- **Voice** (English:2,Spanish:2,Basque:2) Active, Passive.

$$MoprhRatio(D, t_{morph}) = \frac{freq|D, t_{morph}|}{|D_w|} \tag{3.6}$$

where t_{morph} is a tag describing a given morphological phenomena, $|D_w|$ is the total number of words in D and $freq(D, t_{morph})$ the total number of words in D that are associated with t_{morph}.

3.5.3 Syntactic Features

In grammar, syntax is what defines the structure inside a sentence. Therefore, syntactic features are the ones aimed at describing structural complexity. We detail below the syntactic features defined for MultiRead.

PoS Ratios

As described in Section 3.2.6, PoS tags are the building blocks of the structure of a text. Therefore, we hypothesize that some PoS tags are more involved in complex structures than others. In order to analyze this, MultiRead measures the frequency of on occurrence of each PoS tag in a text.

Measuring the occurrence of single PoS tags might not be enough to capture the structure of a text. Therefore, we also measure the occurrences of n-gram PoS tags. As an example, MultiRead measures the frequency of a verb and and adjective appearing one next to the other. As combinations of PoS tags can exponentially grow, we only consider 2-grams and 3-grams. As a result, MultiRead extracts a feature (computed as in Equation 3.7) for each combination of 1, 2 and 3 PoS tags, based on the *frequency of occurrence of that combination* divided by the total number of tokens D.

$$PoSCombinationRatio(D, t_{PoS}) = \frac{freq(D, t_{PoS})}{|D_w|} \tag{3.7}$$

where t_{PoS} denotes one of the aforementioned PoS tag combination, $|D_w|$ is the total number of words in D and $freq(D, t_{PoS})$ the total number of occurrences of t_{PoS} in D.

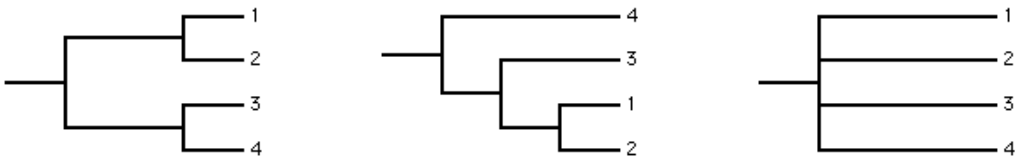
Dependency Tree Complexity

As described in Section 3.2.7, the dependency tree is a deep representation of the structure of a sentence. The assumption is that the more complex this tree is, the more complex is its respective sentence. In order to take advantage of this assumption, MultiRead considers as features the *average complexity of each dependency tree in D*, computed (as in Equation 3.8) based on several complexity metrics. Each metric considered is described below using the tree examples shown in Figure 3.5.

$$DependencyComplexity(D, t_{complexity}) = \sum_{s \in D_s} \frac{complexity(s, t_{complexity})}{|D_s|} \tag{3.8}$$

where $t_{complexity}$ represents a tree complexity method, s is a sentence in D, $|D_s|$ is the total number of sentences in D, and $complexity(s, t_{complexity})$ is the function that yields the level of complexity of a sentence and its corresponding tree, based on $t_{complexity}$.

Figure 3.5: Example trees



(a) Tree 1 (b) Tree 2 (c) Tree 3

- **Depth.** The depth of a tree is the maximum distance from the root to any of its leaves. It is measured counting the nodes that the mentioned maximum path needs to go through. The depth of trees 1, 2 and 3 in Figure 3.5 are 4, 5 and 2 respectively.

- **Resolution.** The resolution of a tree describes how branches occur in a tree. A fully resoluted tree is one that has all branches given the number of nodes (trees 1 and 2 in Figure 3.5), while an unresolved tree is one that has all the leaves directly connected to the root (tree 3 in Figure 3.5) by one branch. For calculating the resolution of a tree, we follow Colless's [38] formula, that simply divides the number of internal branches in a tree by the maximum possible internal branches that the tree could have.

- **Imbalance.** The balance of a tree measures how skewed a tree is to one side.

Tree 1 and 3 in Figure 3.5 are perfectly balanced trees, while tree 2 is the most imbalanced tree possible, since most of the nodes hang from one side of the tree. To measure imbalance of a tree we use Colless's [39] imbalance formula, slightly modified to allow splitting points with more than 2 branches. In this case, $complexity(s, t_{complexity})$ in Equation 3.8 is computed as:

$$Complexity(s, t_{imbalance}) = \sum_{p \in s} \frac{maxSubtree_p - minSubtree_p}{|splittingPoints(s)|} \tag{3.9}$$

where p is a split point in the tree of sentence s, $maxSubtree_p$ is the number of nodes in the biggest subtree generated from p, $minSubtree_p$ is the number of nodes in the smallest subtree generated p and $|splittingPoints(s)|$ is the total number of split points in the tree of s.

- **Branches per splitting point.** Defined as the number of branches that each splitting point has on average. All the example trees have 2.

- **Number of splitting points.** Defined as the number of splitting points in the tree.

  Note that both *Branches per splitting point* and *Number of splitting points* do not describe any specific aspect of the tree, but they are simply oriented to describe the tree from a different perspective and complement the depth, imbalance, and resolution metrics.

3.5.4 Semantic Features

In addition to shallow, syntactic, and morphological features, MultiRead examines semantic features that go beyond the tokens and structure of a text. Doing so

facilitates the analysis of the concepts laying on a text.

Semantic Closeness

The topic of a text is another criteria that can be considered for determining its difficulty. For example, a text about *animals* will generally be simpler to read than a text about *microbiology*. To take advantage of this fact, we created a knowledge base KB which contains information related to the probability of occurrence of topics group by reading levels. In other words, given a pre-defined set of reading levels R and a leveled textual corpora, KB includes the probability distribution of concepts in each reading level $r \in R$, extracted using WordNet. A concept c_i is defined as $c_i = \{w_1, \cdots, w_n\}$, where w_i is a word that refers to the corresponded concept c_i, as determined using WordNet.

Each reading level $r \in R$ is associated with a concept distribution vector $K_r = \{P(c_1), \cdots, P(c_n)\}$, where $P(c_i)$ is the probability of c_i among documents of readability r. Similarly, a vector K_D is created for D, which captures the probability distribution of concepts (defined in KB) in D. Comparing the distribution of concepts in D with respect to the distribution of concepts among documents with different degrees of difficulty, allows MultiRead to identify the most likely reading level of D. This comparison is computed using Equation 3.10, which determines the the similarity between K_D and K_r based on the angle between the two vectors, i.e., cosine similarity strategy [27].

$$SemanticCloseness(D, r) = \frac{K_r \bullet K_d}{||K_r||||K_d||} \tag{3.10}$$

where the numerator is the dot product between K_r and K_d and the denominator the

product between the modules K_r and K_d.

Synonym Usage

As previously stated, the use of synonyms is an important aspect than can determine text quality and therefore readability. Professional writers try to avoid repetition of words by using synonyms and different phrasings, while texts oriented to low reading level audiences tend to repeat more terms to simplify the vocabulary of the text. In order to take advantage of this writing pattern, we use a feature oriented towards measuring the usage of synonyms in a text.

For each concept c_i in KB, we analyze how balanced is the use of each word w_i related to c_i among the all the words in D. The more balanced the frequencies of terms related to a given concept the better use of synonyms by the writer, as repetition has been avoided by using all synonyms equally. On the opposite side, when frequencies are not balanced, i.e., one word has nearly 100% of the occurrences among the terms that pertain to the same concept in D, it means that no effort to use synonyms was made. For measuring the aforementioned balance, we take advantage of Shanon's entropy [97] and create a feature that estimates the *entropy of each synset in* D, as shown in Equation 3.11.

Entropy refers to chaos in an information source. It considers each term as a message and its frequency as the probability to produce that message in order to measure the unpredictability of the signal. The more even the probabilities, the more unpredictable is the signal and therefore the bigger the entropy. In our case, the more even the distribution of terms, the bigger the entropy and therefore the better the synonym usage.

$$SynonymUsage(D) = \sum_{c_i \in D_c} \frac{-\sum_{w_i \in c_i} P(w_i) \log_{10} P(w_i)}{|D_c|} \tag{3.11}$$

where c_i is a concept in D, w_i denotes a term associated with c_i, $P(w_i)$ is the probability of occurrence of w_i in D, D_c is the set of concepts discussed in D and $|D_c|$ is the number of distinct concepts in D.

Cohesion

Cohesion, or coherence, *is the intangible glue that holds paragraphs together* [7]. Having good cohesion in a text means that ideas stick together and that the flow between sentences is smooth. Texts oriented to low skilled readers are usually very cohesive, the jumps from sentence to sentence are small, because the reader has enough trouble understanding each single sentence. More skilled readers precisely comprehend the meaning of each sentence, therefore, they are also more capable of detecting bigger argumentation changes between sentences. In order to capture this phenomena, we measure the *content similarity between adjacent sentences in D*. As described in the following equation, the similarity between sentences is computed using the Jaccard [65] similarity of the sets of non-stopword lemmas in each sentence.

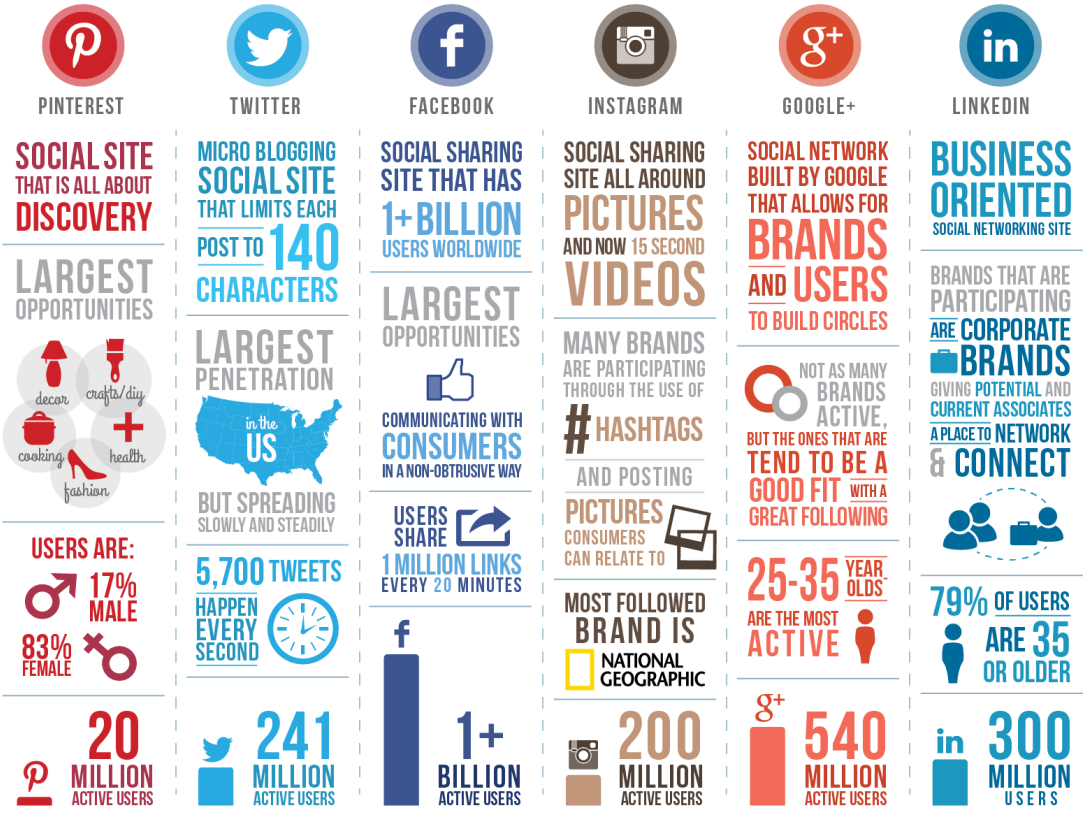
$$Cohesion(D) = \frac{\sum_{s_i, s_j \in D} \frac{|s_i \cap s_j|}{|s_i \cup s_j|}}{|D_{ij}|} \tag{3.12}$$

where s_i and s_j are any two adjacent sentences in D, $|D_{ij}|$ is the number of adjacent sentence pairs in D, $|s_i \cap s_j|$ represents the number of words included in both s_i and s_j and $|s_i \cup s_j|$ captures the number of distinct words in either s_i or s_j.

3.5.5 Social Features

As evidenced by the data captured in Figure 3.6, 5,700 tweets are generated every second, more than 1 billion users use Facebook actively, brands are aware of the importance of their presence in social networks and are actively participating in the using several hashtags.

Figure 3.6: Social network infographic extracted from Leveragemedia.com



Social sites like Facebook, Twitter or Instagram have become very popular, changing the way we see the internet and the resources contained in it. Increasingly, more resources contain social data inherent to these sites, such as hashtags, usernames,

mentions, or links, data that are usually ignored by readability formulas. We define below some of the data that can be found in social media resources:

- **Hashtag.** A hashtag is a term or a set of terms preceded by a # symbol, used to represent a concept in social media. For example, in Figure 3.7 #SIGIR2016 aims to capture that the content of the tweet refers to ACM SIGIR Conference in Information Retrieval.

- **Username.** Most of the documents generated in social media are usually related to a known author. For example, in Figure 3.7 Dr Chole is the name of the user that wrote the tweet.

- **Mention.** A mention is any other user of the social network that is explicitly addressed in a document. Those users are usually directly connected to the author of the document, given a friendship or interest relation. For example, in Figure 3.7 IonMadrazo and NevDragovic are users mentioned in the tweet.

- **Link.** A link is a reference to another document. This document is somehow expected to have some relation with the content of the analyzed document, being a supportive document or a full version of the current document.

- **Emoticon.** An emoticon is a set of punctuation marks oriented to express an emotion such as happiness or sadness. Examples include, :), ;), or :(.

Even if the mentioned information can be important in social media documents, it usually ignored by readability formulas. For this reason, we introduce novel social features we created for MultiRead that take advantage of the aforementioned information.

Figure 3.7: An example Tweet written by Dr Chole.

Frequencies of Social Data

Social media resources contain more information than text, they also contain hashtags, mentions to other users and emoticons. In order to analyze this information for readability prediction we treat the frequencies of occurrence of social media tags, i.e., t_{social}, as predictors for readability. As a result, three features are created for MultiRead, i.e., *ratio of tokens that are hashtags, mentions or emoticons*, each of which is defined as in Equation 3.13.

$$SocialTagRatio(D, t_{social}) = \frac{freq(D, t_{social})}{|D_w|} \tag{3.13}$$

where t_{social} is a given social tag, i.e., hashtag, mention or emoticon, $|D_w|$ is the total number of words in D, and $freq(D, t_{social})$ the number of occurrences of t_{social} in D.

In addition to its general frequency, we also consider the distribution of each individual emoticon, hypothesizing that some emoticons are more adult-friendly than others that are more commonly-used by children. Therefore, for each emoticon, we also consider a feature based on its frequency of occurrence, computed using the following equation:

$$EmoticonRatio(D, t_{emoticon}) = \frac{freq(D, t_{emoticon})}{|D_w|} \tag{3.14}$$

where $t_{emoticon}$ represents a given emoticon, $|D_w|$ is the total number of words in D and $freq(D, t_{emoticon})$ is the total number of occurrences of $t_{emoticon}$ in D.

Extended Information Features

Most of the features described in this manuscript are oriented to be computed over a significant amount of text so that their estimated values are meaningful, i.e., representative of the information they are meant to capture, and serve as evidence to compute accurate readability level predictions. However, social media related documents, such as tweets or comments, only contain a few words, which is usually a restriction. With the goal of incorporating more context that can enhance the quality of the readability prediction for short documents, we also compute all the features described in this manuscript for D', an extended version of D. This extended document includes additional textual information that relates to the content of D. We create four different versions of D', each of which differs depending upon the source of information used to extend D. The sources considered for creating D' include:

- **Other documents written by the user who authored D.** A user is expected to be consistent in his writing in terms of readability. Therefore, when the user name of the writer is known (such as a Facebook username or a twitter handle), it is useful to take advantage of other resources written by him, and use them as context. Based on this information, D' is created by merging all documents written by the author of D.

- **Documents by mentioned users.** Homophily stands for the nature of users to relate to similar users. It is a principle that widely manifests in social networks, in terms of aspects such as age, hobbies or profession of users [112].

Following this principle, our hypothesis is that users with the same readability also tend to relate to each other more frequently that random users. Therefore, we create an extended version of D comprised of documents written by other users mentioned in D, such as their Facebook comments or twitter tweets.

- **Documents that contain same hash-tags.** Following the same principle of homophily, we hypothesize that users with similar readability also share the hashtags they often use. Therefore, we also consider an extended version of D that includes documents that contain the same hashtags as D.

- **Linked web pages.** Web pages linked on a document are usually related to its content. We hypothesize that this relation also exists in term of readability. In other words, it is natural to assume that the readability of a document will be similar to the readability of the resources linked in the document. Therefore, we also create an extended version of D with information extracted from resources linked in D.

3.5.6 Metadata Features

Metadata based features can be useful in environments where text access is limited (i.e., copyrighted material), or the text contains some structure than can influence readability (i.e., webpages). An exploration of this type of feature allows MultiRead to expand the types of texts it can handle. A description of each metadata-based feature is provided below.

Web Page Related Features

Web pages usually contain enough textual context to be assessed by textual readability features. However, the layout and the design of a website also play an important role for predicting the readability of a web-page. A web page full of colors can be oriented towards a younger audience, who usually have low reading skills, while a black and white one can be more professional and therefore more adult oriented. To explicitly account for this fact MultiRead considers the *frequencies of HTML and CSS tags in a web page*, i.e., tags that are responsible for the design and structure of the web page, i.e., D in our case. One feature is created for each web page-related tag, i.e., denoted t_{HTML}, as follows:

$$HTMLRatio(D, t_{HTML}) = \frac{|freq(D, t_{HTML})|}{|D_w|} \tag{3.15}$$

where $|D_w|$ is the total number of words in D, t_{HTML} is an HTML tag and $|freq(D, t_{HTML})|$ is the total number of occurrences of t_{HTML} in D.

Book Related Features

Even if books contain large amounts of textual content, access to their content is usually restricted due to copyright limitations. This makes traditional readability techniques not work on this type of resource. However, as Denning et al. [46] demonstrated, the reading level of a book can be predicted by analyzing metadata about the book that is freely available, even in the absence of sample text.

Following the success reported in [46], MultiRead considers the genre and library of congress subject heading [10] assigned to books, in order to further inform the readability prediction process, in case D is a book for which few sample text is

available to process. While genre refers to "a category of artistic, musical, or literary composition characterized by a particular style, form, or content" [68], e.g., *fiction* or *drama*, subject headings are a " set of keywords used by librarians to categorize and index books according to their themes" [64], e.g., *Trolls, Green* or *the natural language form, e.g., Green Trolls, and the subdivision form, e.g., FantasyMythical CreaturesTrollsGreen."*

Similarly to semantic closeness computed using Equation 3.10, MultiRead also considers the proximity of genre and subject heading distribution of D with respect to the genre and subject headings distributions observed on sample books categorized in R levels of text difficulty. Analyzing the degree of similarity between genres and subject headings assigned to D and those assigned to books for which their level of difficulty is known allows MultiRead to collect further evidence to determine the readability level of D.

3.6 Fusioning Strategy

Individually, the features presented in the previous section can only produce a rough estimate of the reading level of a document, as they only quantify the degree of difficulty of a text from a single perspective. Best results can only be achieved when those features are used in tandem. For doing so, MultiRead takes advantage of Random Forests [34] a machine learning strategy based on multiple decision trees and bagging. The rest of this section describes how this learning model works, starting from background techniques.

3.6.1 Supervised Machine Learning

Supervised machine learning [84] refers to the process of learning a model from labeled data, in order to be able to use it for predicting unknown labels for new data. More formally, each instance x in a collection X is represented as $x :< f_1, ..., f_n, z >$, where f_i is a feature that describes x and z is the known prediction class for x. Using instances in X, a model β is learned that is later applied for predicting the class of any new instance for which z is unknown.

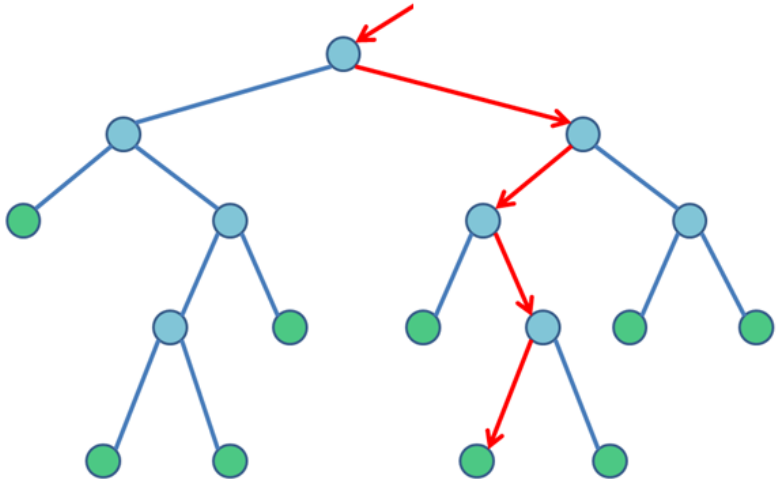
In MultiRead, X corresponds to a labeled corpora of documents for which their corresponding readability level is known. Each document (i.e., x) is associated with a feature representation (based on the features described in Section 3.5) and a predefined degree of text difficulty (i.e., class z). The model β is generated using X and the Random Forest algorithm, which is based on multiple decision trees.

3.6.2 Decision Tree

As previously stated, MultiRead depends upon an algorithm based on decision trees to predict the readability level of D. A decision tree is a structure oriented to make a decision that takes the shape of a tree. Each node of the tree represents a question, and each branch of that node is a possible answer to that question. For example a node can ask *Is the door open?* and each branch has an answer to that *Yes/No*. The decision making process starts at the question at the root and ends in one leaf of the tree. Each leaf has a label which is the result of the whole decision process. Going back to MultiRead each question in a node just asks about the value of a specific feature f and each branch represent a range of possible values, e.g. $f < 1$.

Many machine learning strategies for prediction focus on automatically generating

Figure 3.8: Decision Tree



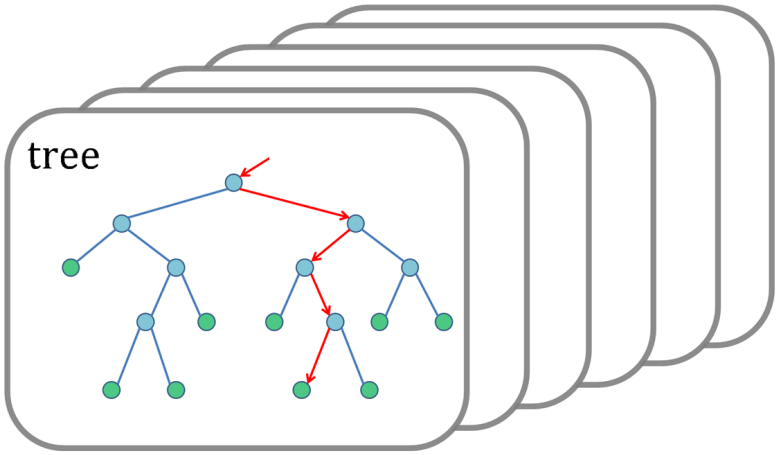
a decision tree over a labeled training dataset, that can later be used to classify any new, unlabeled instance. Among the most popular are C4.5 (J48) [93] and ID3 [92]. These models offer high expressiveness as the tree can grow infinitely creating more branches, potentially being able to learn any combination of data. However, this high expressiveness makes decision trees very prone to overfitting, a common machine learning issue where a model also learns about the noise of a certain dataset and loses its generalization power, failing to predict on new unlabeled data.

3.6.3 Bagging

In order to improve accuracy and avoid overfitting, bagging strategies can be used. Bagging [33] or boostrap aggregating is an ensemble meta-algorithm oriented to improve accuracy of machine learning algorithms by reducing variance and proneness to ovefitting. The idea behind the algorithm is simple, instead of generating a model over all the data available, multiple models over random subsamples of the data are generated. This way each model will not be as overfitted to the entirely of data,

because it has only learned about a portion of it. At the same time, the full dataset will still be used, but the knowledge acquired will be distributed among several models. For generating the final prediction each model will submit a vote, which will be averaged to get the final prediction of the system. As an example, tree bagging (see Figure 3.9) refers to the technique of bagging applied to decision trees, which is the base of Random forests.

Figure 3.9: Example of a tree forest



3.6.4 Random Forest

Random forest refers to a model where decision trees are used within a bagging strategy. As a peculiarity, apart from using a randomized subsample of the data for generating each model, random forests also use a random subsample of the features [75]. Benefits of random forests include high learning power while being less prone than decision trees to overfitting, high accuracy compared to other learning algorithms and efficient scaling, as the model can be easily parallelized given its distributed nature [34]. In addition, the structure of Random Forests match well with the needs

of MultiRead, given the sparse nature of the features employed in MultiRead that can be defined for any given document for prediction purposes. Containing multiple small models, i.e., decision trees, we believe that each of those can specialize in a certain aspect of MultiRead. Some tree may just specialize in tweets for English, while others will be more focused on other languages and document types.

Given the aforementioned arguments and the empirical study presented in the Section 4.4, which verifies the correctness of relying on such a model, we implemented MultiRead so that it incorporates the Random Forest model as its feature fusioning strategy.

CHAPTER 4

EVALUATION

In order to validate the correctness and effectiveness of MultiRead we conducted an in-depth study using several datasets and metrics, which we describe in Sections 4.1 and 4.2. The study is driven by the questions we used to guide our research process: (i) What is the best alternative to fuse evidence (i.e., features) that capture information about documents from different perspectives?, (ii) Which subset of features most influences the readability prediction process?, (iii) Can the performance of a multi-purpose readability tool be consistent among documents of different types, languages, and length?, (iv) Is the tool capable of assessing readability of documents with a more fine-grained degree of difficulty than just simple and complex?, and (v) Are readability formulas/tools consistent when estimating the readability level of the same document in different languages?

Note that, even if MultiRead is designed to be language independent, for practical purposes the results of our evaluation pertain to assessments on corporas in three languages that are representative of the diversity of existing languages: English, Spanish, and Basque, i.e., a germanic, a romance, and a pre-indioeuropean language, respectively.

4.1 Datasets

The ideal dataset for designing, developing, and evaluating MultiRead 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 in English, Spanish, and Basque, that can suit MultiRead's needs and can be used for evaluation purposes. These datasets are summarized in Table 4.1 and described below.

Table 4.1: Overview of the datasets used for validating the design and performance of MultiRead

Dataset	**Sources**	**Languages**	**Document Type**	**Size**	**Number of Target Levels**
BookData	CLDC.com	En	Books	120	12
DMOZData	DMOZ.com	En	Websites	1000	2
Ikasbil	Ikasbil.eus	Eu	Documents	1000	5
TwitterData	Twitter	En	Tweets	22000	6
VikiWiki	Vikipedia.org Wikipedia.org	En, Es, Eu	Documents	12798	2
ParallelData	Albalearning.com	En, Es	Documents	100	1

Vikidia vs. Wikipedia

Wikipedia [13] is an online encyclopedia publicly available for multiple languages. At the same time, Vikidia [11] is a simplified version of the former, containing a sample of the most popular articles of Wikipedia written in a relatively low reading level, so that people with average reading skills, such as children or language learners, can understand them. Using these two sources, we generated a dataset containing all the articles in Vikidia and their corresponding complex counterparts in Wikipedia.

This translated into VikiWiki, a two-level (simple or complex), multilingual (English, Spanish and Basque) dataset comprised of 12798 documents: 1767 documents in English, 4184 in Spanish and 448 in Basque for both simple and complex level.

Twitter Dataset

Currently, there is no dataset available that contains social media resources labeled with their corresponding readability level. Therefore we built one by taking advantage of a Twitter sample of 172M tweets extracted using Twitter's API on 2014 [76, 77]. We followed the simple, yet effective approach presented by Zhang et al. [112] which takes advantage of *happy birthday* messages for determining the age of Twitter users. In doing so, we identified that age of (some of the) Twitter users in our sample, resulting in a dataset of 22k tweets, each of which labeled with the age of the respective author. We are aware that the age of a person is not exactly the same as his expected reading level. Therefore, we split the tweets into six age ranges, following the ranges defined by Levinston [73]. These ranges refer to a person's developmental stages, which given the lack of benchmarks, we consider appropriate to be used as different readability levels. In other words, tweets in this dataset are categorized using *six different levels of text difficulty.*

DMOZ Webpages

DMOZ [8] is the most comprehensive online directory publicly available. It contains links to thousands of categorized webpages. Among these categories, we can identify sites targeting children or mature audiences. Based on these labels,we created a webpage dataset comprised of 500 children-oriented webpages and 500 non-children oriented webpages. As the reading abilities of children are often less developed than

those of an adult, for experimental purposes we considered documents corresponding to children as *simple* documents and remaining ones as *complex*.

Books

We use the book dataset created by Denning et al. [46] using information obtained from the Children's Literature Comprehensive Database Company (CLCD)[6]. This dataset is comprised of 120 books (written in English) that contain book-related information, including title, genre, subject heading and short textual snippets. Furthermore, each book is associated with its readability level in a *K-12* [1] *scale*, which is used as the target readability level for experimental purposes.

Ikasbil

Ikasbil [4] is an online resource for learning Basque that contains different media contents, such as articles, videos, or audio contents. Every resource on the site is labeled given its complexity, using the reading level grading system defined by the Common European Framework of Reference for Languages (CEFR). The dataset contains 200 documents written in Basque for each of the *5 reading levels* defined by CEFR.

Parallel corpus

A parallel corpus is a set of documents that is exactly translated into several languages. This translation needs to be sentence aligned, i.e., each sentence needs to match exactly another sentence in the other languages. We gathered a parallel corpus

[1] K-12 is a term used to refer to primary and secondary grades in American Education System. The system comprises a total of 12 grades, from kindergarten to 12th grade.

from albalearning.com [5], a website oriented to language learning. This corpus is comprised of around 100 documents written in English and Spanish, that are sentence aligned but are not associated with any particular readability label.

4.2 Metrics

To quantify the performance of MultiRead, we use a number of well-known Information Retrieval and Machine Learning metrics often applied to evaluate readability prediction formulas/tools.

Accuracy

Accuracy is used for estimating the performance of a prediction system. It stands for the ratio of correctly classified instances among all instances.

$$Accuracy = \frac{Correctly\ Classified\ Instances}{Total\ Instances} \tag{4.1}$$

where an instance in our case is a test document which is treated as correctly classified only when the predicted readability level matches the known readability level of the document.

Mean Average Error

Mean average error (MAE)[82] refers to the average displacement of the predictions of a system from the true value. It is commonly useful as a substitute of the accuracy metric when prediction values are not binary and instead have some inherent ordinality.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \tag{4.2}$$

where $|f_i - y_i|$ is the absolute error, where f_i is the prediction and y_i the true value and n the number of instances considered.

In our case, n refers to the size of the dataset considered for evaluation purposes, y_i is the known readability level associated with a given instance in the dataset, and f_i is the readability level predicted for the instance using a readability assessment formula/tool.

Cohen's Kappa

Cohen's Kappa [79] measures the agreement among different raters. It is usually used to measure how faithful is an information source provided by various human judges, based on the extent to which they agree with each other. This metric can be also applied to measure agreement among prediction systems.

$$\kappa = 1 - \frac{1 - p_0}{1 - p_e} \tag{4.3}$$

where p_0 is the relative observed agreement among raters and p_e is the hypothetical probability of chance agreement. If there is full agreement $\kappa = 1$, if there is no agreement $\kappa = 0$.

Pearson's Correlation

Pearson's correlation [82] measures how dependent are two variables on each other. It can be used to determine a variable with redundant information or the prediction power of a variable with respect to a class. A positive correlation coefficient between

any two variables means that for every positive increase of one variable, there is a positive increase in the other. A negative coefficient, on the other hand, means that for every positive increase of one variable, there is a negative decrease of the other one. When the value is close to 0 the two variables are not related.

$$r(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt[2]{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt[2]{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{4.4}$$

where X and Y are any two variables, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the average of X; and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ the average of Y.

4.3 Evaluation Strategies

In this Section we present several evaluation strategies oriented to ensure the validity of the results of this study.

N-Fold Cross Validation

N-fold cross validation [82] refers to a technique used to ensure that the assessments obtained for a system using a dataset (regardless of the metric considered) will be generalizable to unknown independent data. This technique splits data into N folds, each of which contains $\frac{1}{N}$ of the data instances. In N different rounds, $N-1$ folds are used for training purposes and 1 fold for testing. The results of each round are then be averaged to obtain a less biased result than what we would obtain just manually splitting the data in one training and one testing set.

Leave-one-out

Leave-one-out [82] is an specific case of N-fold cross validation where the number of folds is equal to the number of instances in the data. This strategy can be useful in environments where the number of data instances is very limited and as much data as possible is needed for training. However, given the number of rounds is the same as the number of instances in the data it can also be computationally expensive in large datasets.

4.4 Which learning model performs better for MultiRead?

Objective The aim of this experiment was to determine which feature fusion strategy fits best for MultiRead, influencing the final decision of which learning model will MultiRead use. For doing so, we analyzed the performance of MultiRead using several learning models. As MultiRead is oriented to multiple document types and languages, the best fusion model for it is the one that consistently performs adequately for all contexts for which MultiRead can be applied.

Dataset To identify the most suitable model, we conducted an empirical study. In order to remove any positive bias that can influence the final evaluation of the system, this study was performed on disjoint datasets, which include 600 documents from the VikiWiki dataset (100 for each language and source), 200 tweets equally distributed among reading levels, 200 webpages from DMOZData also uniformly distributed among reading levels, and 20 books randomly selected from the CLDC dataset. Each of these datasets were used separately for training and testing.

Compared Learning Models We considered several learning models for this experiment. However, we present the most significant ones, i.e., top-3 models in terms of accuracy (Random Forest [34], Decision Tree [93] and Multilayer Perceptron [82]) and the most popular among other readability assessment strategies (Support Vector Machines [103]).

Results Table 4.2 shows the accuracy ratios achieved for each learning model and individual language, as well as the average accuracy obtained by each analyzed model. This accuracy was calculated using a leave-one-out framework over the VikiWi dataset. Random Forest is the model that achieves the best results among the models, regardless of the language considered. These improvements, in terms of accuracy, are statistically significant, as determined using a paired T-Test with a confidence of $p < 0.05$ with respect to the counterparts considered in this study. Support Vector Machines, on the other hand, obtained low accuracy. This is surprising, given that they are one of the most-used tools for readability prediction [29]. A a more in-depth analysis of instances indicated that instances with a lot of missing values were often misclassified by this model, which leads us to determine that data sparsity is the reason for this decrease in accuracy. It is also worth noting the effect the language of the documents have in determining the accuracy of a model. In fact, all models achieve consistently better accuracy ratios for English than for Basque. Given that we used the same features for all the three languages, we can attribute this to two reasons: (1) predicting Basque readability is more complex and (2) the lack of precision of Basque NLP tools makes readability features less precise too, hindering final readability predictions.

We also calculated the accuracy using the aforementioned sampled datasets, which

Table 4.2: Accuracy obtained by MultiRead using different fusion strategies on a disjoint sample of VikiWiki.

	English	**Spanish**	**Basque**	**All**
Random Forest	96.6%	83.1%	81.1%	87.7%
Decision Tree	91.4%	82.9%	79.8%	85.4%
Multilayer Perceptron	84.3%	78.5%	78.4%	81.8%
Support Vector Machines	70.5%	65.1%	62.3%	68.4%

allowed us to quantify and compare the performance of different fusion models, when applied to non-traditional documents, such as tweets, webpages and books. Based on the accuracy ratios reported in Table 4.3, we verified that Random Forests achieve better or comparable performance with respect to those achieved by its counterparts ($p < 0.05$). Therefore, given its consistency among all languages and types of documents we determined that Random Forests is the learning model that best fits MultiRead.

Table 4.3: Accuracy obtained by MultiRead using different fusion strategies to predict the readability level of documents of different type.

	Long Snippets	**Social**	**Webpages**	**Books**
Random Forest	87.7%	83.1%	79.2%	45.1%
Decision Tree	85.4%	82.9%	77.6%	40.4%
Multilayer perceptron	81.8%	80.1%	79.3%	46.7%
Support Vector Machines	68.4%	81.3%	75.9%	44.2%

4.5 Which feature subset performs better for MultiRead?

Objective As described in previous sections, a good feature set is what determines the effectiveness of a machine learning-based system, and therefore what will determine the overall effectiveness of MultiRead. The objective of this experiment is to

extensively analyze the influence each feature (or group of features) incorporated in MultiRead has in predicting the reading level of a document.

Dataset The dataset used for this experiment is comprised of 3000 documents from the VikiWiki dataset equally distributed among languages (English, Spanish and Basque) and source (Vikidia and Wikipedia). Even if this dataset has more than 3000 documents, we discarded some documents in English and Spanish in order to evenly balance the distribution among languages in the dataset. We also included 1000 tweets, 500 webpages and 120 books as separate datasets. Note that this dataset is completely disjoint from the one used in 4.4 to remove any bias. It is also important to note that we converted all readability levels into a binary range, i.e., simple or complex, splitting the readability levels in two halves (the most complex half and the most simple half) for those document types for which the class was not binary.

Table 4.4: Accuracy obtained by MultiRead using each group of features.

	Shallow	**Morphological**	**Syntactic**	**Semantic**	**Social**	**Metadata**
English	75.4%	52.2%	76.3%	60.8%	60.9%	53.3%
Spanish	70.6%	64.8%	78.2%	61.3%	NA	NA
Basque	55.4%	65.2%	80.3%	62.0%	NA	NA
Overall	63.1%	57.1%	78.5%	61.2%	60.8%	53.3%

Result In order to measure the predictive importance of each feature, we grouped them given their linguistic category and trained MultiRead using only one feature set at a time. We computed the accuracy of the model generated by each feature subset for each individual language using 10-cross fold validation. As presented in Table 4.4, syntactic features are the ones with most predictive power, followed by shallow and semantic features. This was expected, given that the superiority of syntactic

features was also reported by other readability studies [45, 67], demonstrating their robustness in both short and long documents. Morphological features have one of the lowest accuracy overall. However, looking into more detail, we noticed that this low accuracy is caused by their low performance for the English language, as the accuracy value for this groups of features for English is significantly lower than for Spanish and Basque. This result is expected, as English morphology is not as rich as Spanish's or Basque's morphology. A similar pattern was observed for shallow features, where the accuracy for Basque is significantly lower than the one achieved by the two other languages, demonstrating that shallow features are not as successful for this language. Semantic features are the most consistent among all languages. We hypothesize that this is due to the use of concepts, that are language independent, in contrast to syntactic or shallow features that can depend on the grammar and vocabulary of each language. Social and metadata features, only available for English due to dataset constraints, have some predictive power on their own, which validates our hypothesis: for documents with short textual information, considering data inherent from the type of the document has a positive influence in terms of correctly predicting the readability of the document.

In order to perform a deeper analysis, apart from the feature groups, we also performed the same study for each individual feature, identifying the top 10 features that most influence the readability prediction process for each language. For doing so, we calculated the correlation for each feature used in MultiRead. The lists of top-10 for each English, Spanish and Basque are described in Table 4.5, Table 4.6, and Table 4.7.

Based on the correlation numbers presented in the tables, we further demonstrate that it is harder to predict the readability level for texts in Basque than text in

English: the correlations reported for the top-10 ranked features for Basque are half of the Pearson's correlation reported for the corresponding features for English.

Further analysis of the features indicate that shallow features are prominent among the highest-correlated features for English, as *the unique term ratio* and the *average sentence length* are both among the top-3 features. *Unique term ratio* is the highest-correlated feature for Spanish, with a slightly lower correlation. This feature, however, does not even appear in top-10 features for Basque. We argue that obtaining unique terms for Basque is less precise, as Basque lemmatization is more complex, and thus this feature is not a strong indicator for readability prediction in Basque.

As previously illustrated in Table 4.4, morphology is not an important aspect to consider for predicting the readability of texts in English. This is further verified based on the fact that no morphological features appear among the top-10 features for English, while four morphological features are shown among the top-10 for Spanish and two among the top-10 features for Basque. We also observed that connectors are very influential for Basque readability prediction, as six features out of ten reported in Table 4.7 are based on connectors.

Table 4.5: Top 10 most influential features in terms of readability prediction with their correlation values for English.

Correlation	**Feature**
.70	(Shal) Unique Term Ratio
.62	(Shal) Period Ratio
.55	(Shal) Sentence Length
.47	(Syn) Probability of a noun modifier with a direct object as child
.46	(Syn) Probability of an adverb from the root node
.43	(Sem) Cohesion
.42	(Syn) Probability of a subject from the root node
.41	(Syn) Probability of an object from the root node
.41	(Shall) Ratio of opening punctuations
.40	(Syn) Probability of a modifier from the root node

Table 4.6: Top 10 most influential features in terms of readability prediction with their correlation values for Spanish.

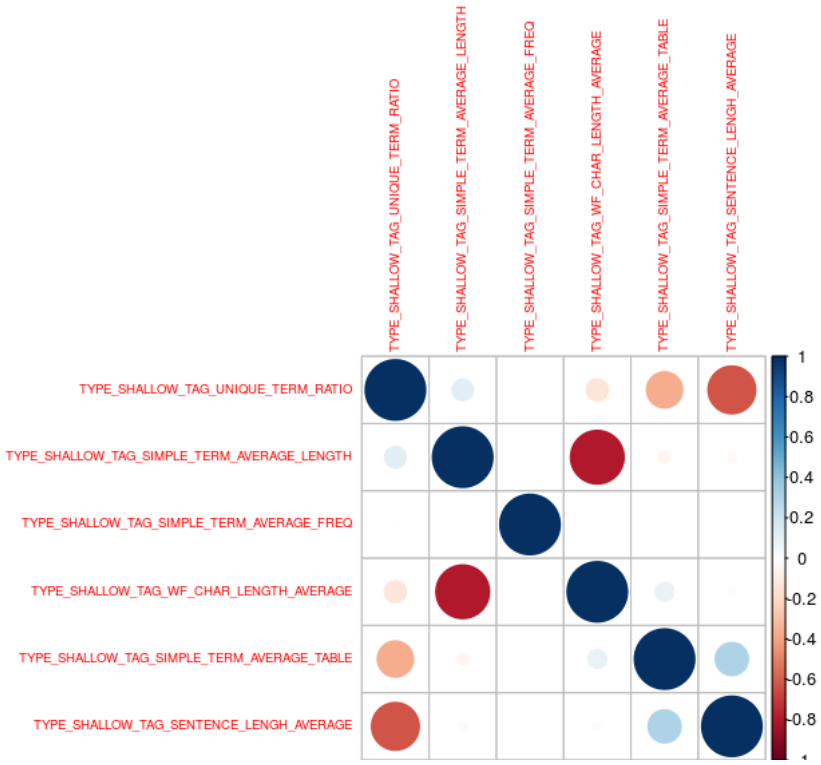
Correlation	**Feature**
.59	(Shal) Unique Term ratio
.48	(Shal) Period Ratio
.44	(Morph) Ratio of present tense
.42	(Morph) Probability of semiauxiliars
.42	(Morph) Probability of indicative mood
.37	(Syn) Probability of a noun followed by a punctuation
.37	(Shal) Sentence Length
.37	(Syn) Probability of a puntuctuation from the root node
.34	(Syn) Probability of a determinant followed by a noun
.31	(Morph) Probability of indefinite determiners

Table 4.7: Top 10 most influential features in terms of readability prediction with their correlation values for Basque.

Correlation	**Feature**
.32	(Syn) Probability of a punctuation followed by a connector
.28	(Syn) Probability of a modifier followed by a punctuation
.32	(Syn) Probability of a connector followed by a punctuation
.27	(Morph) Probability of present tense
.24	(Morph) Probability of indicative mood
.23	(Syn) Probability of a connector with a punctuation as a child
.23	(Syn) Probability of a connector from the root node
.22	(Syn) Probability of an adverb
.21	(Syn) Probability of a connector that has a modifier as a child
.21	(Syn) Probability of a connector that has another connector as a child

Correlation with the reading level, however, is not the only important aspect for a feature. MultiRead considers around 11k features, most of them being 0 or non applicable in most of the documents, generating a very sparse set of features for each document. In order to amend this issue, it is also important to have redundancy, in terms information captured by one or more features examined by MultiRead. This

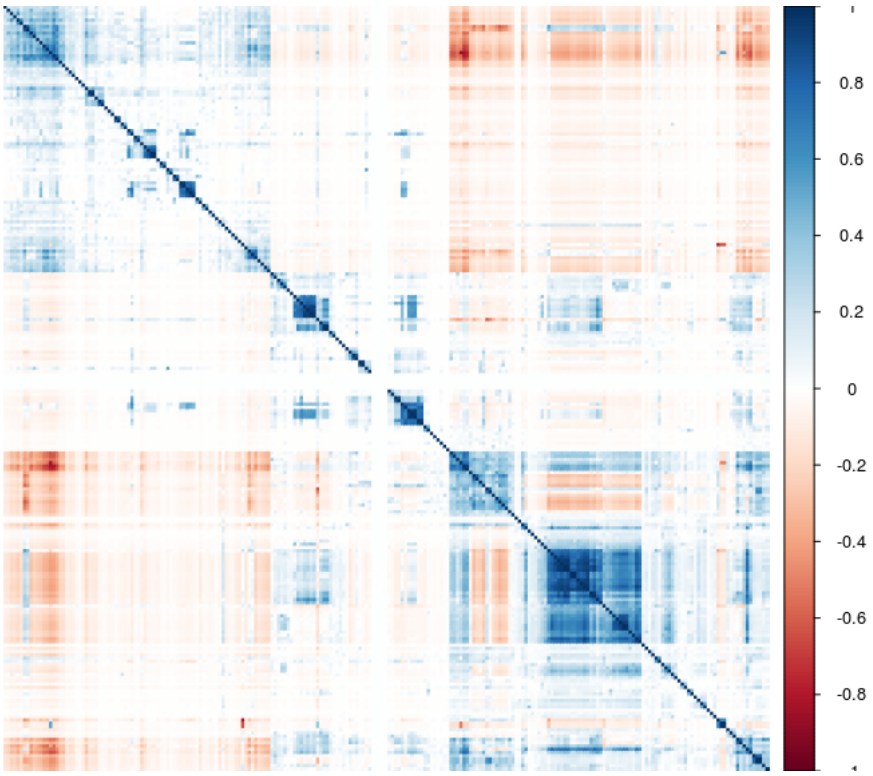
Figure 4.1: Correlations of shallow features



way, if some features cannot be computed for a document, other similar features can replace them to still be able to produce accurate readability level predictions.

The correlations between all shallow features can be seen in Figure 4.1. The figure describes a high correlation between the average term length and the ratio of simple terms, as well as average sentence length and unique term ratio. This high correlation, implies that when any one the features of the pairs is missing or fails for some reason, the other can be used as replacement, ensuring the consistency and robustness of MultiRead. A similar effect can be observed in Figure 4.2, where the correlation for 250 random features is depicted. For visualization purposes, features have been clustered given their correlation, generating several groups of highly correlated features. Those clusters take the form of big blue squares, where

Figure 4.2: Correlations of 250 random features



each of the features in the square is highly correlated with all other features in it. We can see several clusters in the figure, demonstrating that there are indeed group of features that can be replace with others, in terms of capturing necessary information for readability prediction purposes.

4.6 How does MultiRead perform? (overall, by language, by document type)

Objective So far, we analyzed MultiRead's performance in terms of the models and features it considers. However, it is imperative to evaluate its overall performance and compare it with that obtained by other readability prediction tools, which is the

focus of this experiment.

Compared Strategies We compare MultiRead with 3 traditional readability formulas, commonly used by institutions and schools for determining the readability of a text, i.e., Flesch [54], Dale-Chall [42] and Smog [81] readability formulas. Note that we also planned to compare MultiRead with the only other system developed for more than one language, which is the system presented by Clercq et al. [44] (See Section 2 for further details). Unfortunately, obtaining the dataset for that system was not possible.

Dataset For this experiment we used the same dataset as in Section 4.5, comprised of 3000 WikipediaVsVikidia articles, 1000 tweets, 500 webpages and 120 books.

Table 4.8: Comparison of accuracy for different readability assessment strategies.

	Flesch	**Dale-Chall**	**SMOG**	**MultiRead**
WikipediaVsVikidia (English)	70%	72%	69%	96%
WikipediaVsVikidia (Spanish)	65%	68%	66%	83%
WikipediaVsVikidia (Basque)	57%	59%	57%	81%
Tweets	28%	30%	29%	80%
Webpages	65%	57%	67%	82%
Books	23%	27%	21%	53%

Result As shown in Table 4.8, MultiRead outperforms the other 3 readability assessment strategies in all the cases. The difference is specially significant when estimating the readability levels of tweets, books and documents in the Basque version of the VikiWiki dataset. This difference demonstrates the value of both social and metadata features, as well as some morphological and syntactical features specifically designed for Basque. Books accuracy, however, is low compared to other document

types, which is natural, given the lack of text in these documents and the number of possible readability levels (i.e., 12 level), which makes the prediction task more difficult than when only two levels are considered (i.e., simple and complex).

4.7 How well does MultiRead perform when the prediction of readability goes beyond binary levels?

Objective All the experiments performed so far evaluated the effectiveness of MultiRead using binary readability values, i.e., simple or complex. The reason for this lies in the lack of large datasets labeled for more than two readability levels. To validate that MultiRead is indeed multipurpose and can also work in a non-binary environment, we conducted an initial experiment based on one of the few multilevel datasets available.

Dataset For this experiment, we took advantage of the Ikasbil dataset. This dataset is comprised of 1000 documents written in Basque and uniformly distributed among 5 readability levels (i.e., 200 documents per level).

Compared Strategies As we have done for previous experiments, we compare our results with those obtained by Flesch, Dale-Chall and Smog for for the same dataset.

Table 4.9: The performance of MultiRead on the the Ikasbil multilevel dataset, in terms of accuracy and mean average error.

	Flesch	**Dale-Chall**	**Smog**	**MultiRead**
Mean Average Error	1.8	1.9	1.8	0.7
Accuracy	27%	25%	31%	62%

Result As shown in Table 4.9, MultiRead outperforms the readability assessment strategies considered in this experiment in terms of both accuracy and mean average error. An accuracy of 62% might may seem low compared to the accuracy ratios reported for Basque in previous experiments. This is due to a higher number of possible readability values for prediction. However, a deeper exploration of the misclassified instances revealed that the error in classification is, on average, 0.7, as reported by computing the Mean Average Error measure on the same dataset. This low error rate demonstrates that not only the discrepancy in readability prediction error is of less than one readability level on the average, but also this error is justifiable in readability assessment as even human experts have discrepancy when determining the reading level of a text [56].

4.8 Are readability predictions of MultiRead the same for different languages?

Objective As previously mentioned, one of the benefits of multilingual readability assessment is the possibility to use it for ensuring correct translation of documents, as translators can verify that the readability level of the translated document is faithful for the one of the original documents. For this, we need to make sure that the readability predictions for a same text translated to different languages are similar. Therefore in this experiment we measure the agreement of MultiRead's readability predictions among different languages and compare it to other readability assessment strategies.

Dataset Given the lack of multilevel multilingual datasets publicly available, we use two different datasets for this experiment. First, we use WikipediaVsVikidia to train MultiRead, using 3000 texts, 500 per language and reading level. In addition, we take advantage of a parallel corpus in English and Spanish to conduct this experiment. Note that the intent of this experiment is only to assess the agreement of MultiRead predictions, but not their accuracy. Therefore we do not require to know the reading level of each document in the parallel corpus, as we just need to know that the documents are exact translations of each other.

Table 4.10: Inter Language agreement of readability predictions.

	Flesch	**Dale-Chall**	**Smog**	**MultiRead**
Ratio of agreement	57%	61%	63%	84%
Kappa	0.19	0.21	0.22	0.69

Result We calculated the readability for each sentence and their corresponding translation in the parallel corpus. Table 4.10 shows the agreement on the prediction of each pair of sentences, for different readability assessment strategies. MultiRead shows the highest agreement among the predictions of readability for each sentences pair, outperforming other strategies by 20% of agreement. This is translated into a Cohen's Kappa value of 0.69 which describes a *Good* agreement, while the values for other strategies show a *Poor* agreement based on Altman's categorization [21].

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

In this manuscript we introduced MultiRead, a multipurpose readability assessment system pioneer in its area. MultiRead is capable of handling documents of multiple languages, format and length. MultiRead takes advantage of multiple textual patterns that portray a text from different perspectives and fusions this information using Random Forests, in order to predict the readability level of the text.

We explored and designed multiple features based on different textual information: *shallow features*, that examine basic textual information, such as sentence length, *morphological features*, that capture information pertaining to how words are formed, *syntactic features*, that analyze the structure of sentences, *semantic features*, that go beyond words in a text to take advantage of concepts, and *social and metadata features*, that serve in lieu of textual content, when text is practically non-existent or the document type is non-traditional (e.g., a tweet). Social-related features, specially those referring to the data inherent of stylings and wtitting patterns observed on social media sites, which are increasingly more prominent nowadays, are a major contribution of our word as these features based on hashtags, shorthand notation, and emoticons, are usually overlooked by traditional readability assessment formulas/tools.

We performed an in-depth study to validate the correctness of MultiRead and

acquire in-depth knowledge about the data and tools common among Language, Information Retrieval, Machine Learning, and Natural Language Processing areas of study, which gave us the opportunity to understand the benefits and constraints of multipurpose readability assessment systems, such as the one discussed in this thesis. The study revealed MultiRead is capable of predicting the readability of a document regardless of its language or type, outperforming other commonly used readability prediction strategies such as Flesch or Dale-Chall. In addition, we demonstrated that the novel social features we introduced are of great value for readability level prediction of social documents, such as tweets, that are of special difficulty for traditional readability assessment formulas/tools given their short length and non-traditional use of text. Finally, we also proved the agreement of MultiRead predictions in documents translated to other languages, making the tool usable for applications that need readability assessment for more that one language at a time.

During the research process that lead to MultiRead we had to overcome several issues, such as the lack of multilingual multidocument readability labeled datasets. For doing so we explored several textual resources that together, permitted the assessment of the performance of MultiRead and other tools. This textual resources will be made available to the community as a byproduct of this thesis.

We anticipate that the availability of a multipurpose readability assessment tool such as MultiRead will be beneficial for a number of applications. The most obvious one is the readability prediction of single documents, which now will be possible using a unique tool that yields reading level predictions in a unified scale regardless of the language or the document type. *Educators*, *public institutions* and *librarians* will be able to use a single tool regardless of the text they need to analyze. In addition, educational applications that make use of multiple readability formulas will no longer

need to integrate multiple tools, facilitating the process of integration and ensuring agreement of readability predictions among all languages.

Professional books translators, will also benefit from MultiRead as they will be able to ensure that their translations are not only correct in terms of content but that they are also faithful to the intended reading level of the original documents.

Social applications that previously ignored the reading level of documents such as hashtag recommendation (see Appendix A), user recommendation, advertisement targeting, re-tweet prediction or search engines oriented to people with diverse reading abilities [25, 40], will now be able to use MultiRead to improve their performance.

More importantly, any single application dealing with *non-traditional users*, such as children, language learners or people with reading difficulties, will be able to integrate readability assessment in tasks such as retrieval of documents, recommendation or other personalization tasks.

We are aware that there are limitations that still need to be addressed in MultiRead. While they are beyond the scope of this project, they open new research avenues that we would like to explore in the future.

The biggest constraint we encountered during our research process was the shortage of benchmarks. As MultiRead is the first system of its type in the area, we expected that lack of datasets that contained multilingual documents of different types labeled according their level of (text) difficulty. However, we did not anticipate the lack of single language readability assessment benchmarks. There is not official benchmark for readability assessment that permits the unbiased comparison of results among other tools that perform the same task. To solve this issue, we plan to develop a multilingual readability assessment benchmark comprised of multiple types of texts using human judgments. This would create an unified benchmark, authors could

use to assess the performance of their system and compare it to others in the area. However this is not trivial and will require research and professional effort.

Using the aforementioned dataset, we would also like to perform a more in-depth study that considered more languages beyond English, Spanish and Basque, as well as expand our assessment in non-binary readability prediction.

Finding text processing tools that fitted MultiRead's needs was not a trivial task either. SyntaxNet could have been the perfect tool for developing it. However, given that it still does not offer some basic features we decided to discard it from MultiRead for now. However, this tool might be of interest in the near future. Once SyntaxNet is fully developed, it could be included in MRAS, easily extending MRAS language compatibility to over 70 languages.

Regarding prediction features, we would also like to explore features that take the pragmatics of a text into account. Pragmatics study how the general structure of a text is organized, a fact that we think could be of interest for readability prediction. We would also like to do more research in more semantic features, as they showed to be the most consistent among all languages. We would like to use a more precise technique for extracting concepts than WordNet and take advantage it for building new features. We are also aware about the relatively poor accuracy provided by book related features and we plan to do further research in this area.

Even if there is still a long way towards multipurpose readability assessment, we believe that we established a precedent with the development of MultiRead that will shape the research future of the readability assessment area.

Analyzing... (97%)

Analyzing... (98%)

Analyzing... (99%)

Analysis finished.

Readability of "IonMadrazo_ThesisReport.pdf": 91 (Graduate Level)

REFERENCES

[1] http://opennlp.apache.org.

[2] https://github.com/tensorflow/models/tree/master/syntaxnet.

[3] http://universaldependencies.org/.

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

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

[6] Children's literature comprehensive database. http://www.clcd.com/#/advancedsearch.

[7] Cohesion. http://www.clarkson.edu/writingcenter/docs/cohesion.pdf/. Greg Dorchies, Clarkson University.

[8] Dmoz. https://www.dmoz.org/.

[9] Stopword lists for 50 languages. https://github.com/6/stopwords-json. 2016 Peter Graham.

[10] Subject headings library of congress. http://id.loc.gov/authorities/subjects.html.

[11] Vikidia. http://www.vikidia.org/.

[12] Who are our publisher partners. lexile. https://www.lexile.com/about-lexile/how-to-get-lexile-measures/text-measure/publishers/who-else-is-doing-it/.

[13] Wikipedia. http://www.wikipedia.org.

[14] Itziar Aduriz, Maxux J Aranzabe, Jose Maria Arriola, Arantza Díaz de Ilarraza, Koldo Gojenola, Maite Oronoz, and Larraitz Uria. A cascaded syntactic analyser for basque. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 124–134. Springer, 2004.

[15] Aitor González Agirre, Egoitz Laparra, German Rigau, and Basque Country Donostia. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *GWC 2012 6th International Global Wordnet Conference*, page 118, 2012.

[16] Amani A Al-Ajlan, Hend 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.

[17] Judith Albright, Carol de Guzman, Patrick Acebo, Dorothy Paiva, Mary Faulkner, and Janice Swanson. Readability of patient education materials: implications for clinical practice. *Applied Nursing Research*, 9(3):139–143, 1996.

[18] J Charles Alderson. *Reading in a foreign language: A reading problem or a language problem*. 1984.

[19] Inaki Alegria, Olatz Arregi, Irene Balza, Nerea Ezeiza, Izaskun Fernandez, and Ruben Urizar. Design and development of a named entity recognizer for an agglutinative language. In *First International Joint Conference on NLP (IJCNLP-04). Workshop on Named Entity Recognition*, 2004.

[20] Iñaki Alegria, Xabier Artola, Kepa Sarasola, and Miriam Urkia. Automatic morphological analysis of basque. *Literary and Linguistic Computing*, 11(4):193–203, 1996.

[21] DG Altman. Inter-rater agreement. *Practical statistics for medical research*, 5:403–409, 1991.

[22] Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina 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.

[23] Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. Globally normalized transition-based neural networks. *arXiv preprint arXiv:1603.06042*, 2016.

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

[25] Ion Madrazo Azpiazu, Nevena Dragovic, and Maria Soledad Pera. Finding, understanding and learning: Making information discovery tasks useful for children and teachers. In *SAL 2016-Proceedings of the 2nd International*

Workshop on Search as Learning, co-located with the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2016, 2016.

[26] Ion Madrazo Azpiazu and Maria Soledad Pera. Is readability a valuable signal for hashtag recommendations? *ACM RecSys 2016*.

[27] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[28] Kepa Bengoetxea and Koldo Gojenola. Application of different techniques to dependency parsing of basque. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 31–39. Association for Computational Linguistics, 2010.

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

[30] Gretchen K Berland, Marc N Elliott, Leo S Morales, Jeffrey I Algazy, Richard L Kravitz, Michael S Broder, David E Kanouse, Jorge A Muñoz, Juan-Antonio Puyol, Marielena Lara, et al. Health information on the internet: accessibility, quality, and readability in english and spanish. *Jama*, 285(20):2612–2621, 2001.

[31] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python*. " O'Reilly Media, Inc.", 2009.

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

[33] L Breiman. Bagging predictors (technical report). *University of California, Department of Statistics*, 1994.

[34] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[35] Jeanne Sternlicht Chall and Edgar Dale. *Readability revisited: The new Dale-Chall Readability Formula*. Brookline Books, 1995.

[36] Jean-Charles Chebat, Claire Gelinas-Chebat, Sabrina Hombourger, and Arch G Woodside. Testing consumers' motivation and linguistic ability as moderators of advertising readability. *Psychology & Marketing*, 20(7):599–624, 2003.

[37] Yaw-Huei Chen, Yi-Han Tsai, and Yu-Ta 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.

[38] DH Colless. Congruence between morphometric and allozyme data for menidia species: a reappraisal. *Systematic Zoology*, 29(3):288–299, 1980.

[39] Donald H Colless. Review of phylogenetics: the theory and practice of phylogenetic systematics. *Syst. Zool*, 31:100–104, 1982.

[40] Kevyn Collins-Thompson, Paul N Bennett, Ryen W White, Sebastian De La Chica, and David Sontag. Personalizing web search results by reading level. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 403–412. ACM, 2011.

[41] Kevyn Collins-Thompson and James P Callan. A language modeling approach to predicting reading difficulty. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 193–200, 2004.

[42] Chall J Dale E. A formula for predicting readability. 1948.

[43] Alice Davison and Robert N Kantor. On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*, pages 187–209, 1982.

[44] Orphée De Clercq and Veronique Hoste. All mixed up? finding the optimal feature set for general readability prediction and its application to english and dutch. *Computational Linguistics*, 2016.

[45] Felice Dell'Orletta, Simonetta Montemagni, and Giulia 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.

[46] Joel Denning, Maria Soledad Pera, and Yiu-Kai Ng. A readability level prediction tool for k-12 books. *Journal of the Association for Information Science and Technology*, 67(3):550–565, 2016.

[47] Biljana Drndarević, Sanja Štajner, Stefan Bott, Susana Bautista, and Horacio Saggion. Automatic text simplification in spanish: a comparative evaluation of complementing modules. In *Computational Linguistics and Intelligent Text Processing*, pages 488–500. Springer, 2013.

[48] Mahmoud El-Haj and Paul Edward Rayson. Osman–a novel arabic readability metric. 2016.

[49] Hillary Evans, Morgan G Chao, Cortney M Leone, Michael Finney, and Angela Fraser. Content analysis of web-based norovirus education materials targeting consumers who handle food: An assessment of alignment and readability. *Food Control*, 65:32–36, 2016.

[50] Bin Fang, Qiang Ye, Deniz Kucukusta, and Rob Law. Analysis of the perceived value of online tourism reviews: influence of readability and reviewer characteristics. *Tourism Management*, 52:498–506, 2016.

[51] Lijun Feng. Automatic readability assessment for people with intellectual disabilities. *ACM Special Interest Group on Accessible Computing*, (93):84–91, 2009.

[52] Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie 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.

[53] Rudolph Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221, 1948.

[54] Rudolph Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221, 1948.

[55] Jonathan Neil Forsyth. *Automatic Readability Prediction for Modern Standard Arabic*. PhD thesis, Brigham Young University, 2014.

[56] Thomas François and Cédrick 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.

[57] Daniela B Friedman, Laurie Hoffman-Goetz, and Jose F Arocha. Health literacy and the world wide web: comparing the readability of leading incident cancers on the internet. *Medical informatics and the Internet in medicine*, 31(1):67–87, 2006.

[58] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics, 2011.

[59] Fréderic Godin, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwen, and Rik Van de Walle. Using topic models for twitter hashtag recommendation. In *WWW*, pages 593–596. ACM, 2013.

[60] Itziar Gonzalez-Dios, Marıa Jesús Aranzabe, Arantza Dıaz de Ilarraza, and Haritz Salaberri. Simple or complex? assessing the readability of basque texts. In *Proceedings of International Conference on Computational Linguistics*, volume 2014, 2014.

[61] Robert Gunning. {The Technique of Clear Writing}. 1952.

[62] Sandra Hale and Stuart Campbell. The interaction between text difficulty and translation accuracy. *Babel*, 48(1):14–33, 2002.

[63] Ahmed MS Ibrahim, Christina R Vargas, Pieter GL Koolen, Danielle J Chuang, Samuel J Lin, and Bernard T Lee. Readability of online patient resources for melanoma. *Melanoma Research*, 26(1):58–65, 2016.

[64] Sheila S Intner, Joanna F Fountain, and Jean Riddle Weihs. *Cataloging correctly for kids: An introduction to the tools*. American Library Association, 2010.

[65] Paul Jaccard. *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz, 1901.

[66] Tapas Kanungo and David Orr. Predicting the readability of short web summaries. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 202–211. ACM, 2009.

[67] Nikolay Karpov, Julia Baranova, and Fedor Vitugin. Single-sentence readability prediction in russian. In *International Conference on Analysis of Images, Social Networks and Texts_x000D_*, pages 91–100. Springer, 2014.

[68] Brett Kessler, Geoffrey Numberg, and Hinrich Schütze. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 32–38. Association for Computational Linguistics, 1997.

[69] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document, 1975.

[70] George R Klare. A second look at the validityl of readability formulas. *Journal of Literacy Research*, 8(2):129–152, 1976.

[71] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Chris Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 2014.

[72] Colleen Lennon and Hal Burdick. The lexile framework as an approach for reading measurement and success. *Electronic publication on https://cdn.lexile.com/m/resources/materials/Lennon__Burdick_2004.pdf*, 2004.

[73] Daniel J Levinson. A conception of adult development. *American psychologist*, 41(1):3, 1986.

[74] Daniel J Levinson. A conception of adult development. *American psychologist*, 41(1):3, 1986.

[75] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[76] Axel Magnuson. *Evaluation of Topic Models for Content-Based Popularity Prediction on Social Microblogs*. Boise State University, 2016.

[77] Deepa Mallela. *CEST: City Event Summarization using Twitter*. Boise State University, 2016.

[78] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.

[79] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.

[80] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60, 2014.

[81] G Harry Mc Laughlin. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646, 1969.

[82] Ryszard S Michalski, Jaime G Carbonell, and Tom M Mitchell. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

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

[84] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

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

[86] Llus Padr, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene 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.

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

[88] Chirag R Patel, Saurin Sanghvi, Deepa V Cherla, Soly Baredes, and Jean Anderson Eloy. Readability assessment of internet-based patient education materials related to parathyroid surgery. *Annals of Otology, Rhinology & Laryngology*, pages 523–527, 2015.

[89] Maria Soledad Pera and Yiu-Kai 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.

[90] Jennifer Petkovic, Jonathan Epstein, Rachelle Buchbinder, Vivian Welch, Tamara Rader, Anne Lyddiatt, Rosemary Clerehan, Robin Christensen, Annelies Boonen, Niti 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.

[91] Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195. Association for Computational Linguistics, 2008.

[92] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[93] J Ross Quinlan. C4. 5: Programming for machine learning. *Morgan Kauffmann*, page 38, 1993.

[94] Richard D Robinson, Michael C McKenna, and Judy M Wedman. Issues and trends in literacy education. 2000.

[95] Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):14, 2015.

[96] Sarah E Schwarm and Mari Ostendorf. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 523–530. Association for Computational Linguistics, 2005.

[97] CE Shannon. A mathematical theory of communication, bell system technical journal 27: 379-423 and 623–656. *Mathematical Reviews (MathSciNet): MR10, 133e*, 1948.

[98] Sally E Shaywitz, Michael D Escobar, Bennett A Shaywitz, Jack M Fletcher, and Robert Makuch. Evidence that dyslexia may represent the lower tail of a normal distribution of reading ability. *New England Journal of Medicine*, 326(3):145–150, 1992.

[99] George Spache. A new readability formula for primary-grade reading materials. *The Elementary School Journal*, 53(7):410–413, 1953.

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

[101] Sanja Štajner, Ruslan Mitkov, and Gloria Corpas Pastor. Simple or not simple? a readability question. In *Language Production, Cognition, and the Lexicon*, pages 379–398. Springer, 2015.

[102] Sanja Štajner and Horacio 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.

[103] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[104] Liling Tan and Francis Bond. Building and annotating the linguistically diverse ntu-mc (ntu-multilingual corpus). *Int. J. of Asian Lang. Proc.*, 22(4):161–174, 2012.

[105] Kumiko Tanaka-Ishii, Satoshi Tezuka, and Hiroshi Terada. Sorting texts by readability. *Computational Linguistics*, 36(2):203–227, 2010.

[106] A Uitdenbogerd. Readability of french as a foreign language and its uses. In *ADCS 2005: The Tenth Australasian Document Computing Symposium*, pages 19–25. University of Sydney, 2005.

[107] Hao Xing Wang. Developing and testing readability measurements for second language learners. 2016.

[108] Barry D Weiss, Mary Z Mays, William Martz, Kelley Merriam Castro, Darren A DeWalt, Michael P Pignone, Joy Mockbee, and Frank A Hale. Quick assessment of literacy in primary care: the newest vital sign. *The Annals of Family Medicine*, 3(6):514–522, 2005.

[109] Chen-Hsiang Yu and Robert C Miller. Enhancing web page readability for non-native readers. In *Proceedings of the sIGCHI conference on human factors in computing systems*, pages 2523–2532. ACM, 2010.

[110] Eva Zangerle, Wolfgang Gassler, and Gunther Specht. Recommending #-tags in twitter. In *SASWeb 2011*, volume 730, pages 67–78, 2011.

[111] Jinxue Zhang, Xia Hu, Yanchao Zhang, and Huan Liu. Your age is no secret: Inferring microbloggers' ages via content and interaction analysis. In *AAAI ICWSM*, 2016.

[112] Jinxue Zhang, Xia Hu, Yanchao Zhang, and Huan Liu. Your age is no secret: Inferring microbloggers ages via content and interaction analysis. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

APPENDIX A

IS READABILITY A VALUABLE SIGNAL FOR HASHTAG RECOMMENDATIONS?

PUBLISHED AT ACM RECSYS 2016

We present an initial study examining the benefits of incorporating readability indicators in social network-related tasks. In order to do so, we introduce TweetRead, a readability assessment tool specifically designed for Twitter and use it to inform the hashtag prediction process, highlighting the importance of a readability signal in recommendation tasks.

A.1 Introduction

Readability is a measure of the ease with which a text can be read. Usually represented by a number, it is an indicator used by teachers to classify and find appropriate resources for students. Several studies have demonstrated the benefits of using readability indicators in educational-related applications, such as book recommendation, text simplification, or automatic translation. However, applying readability indicators outside this environment remains relatively unexplored. Social networks could benefit from readability assessment. Twitter is a social network where users and texts are the main focus. For this reason, it is natural to think that for Twitter the ease with

which a tweet can be understood by a user may affect his interest in it, and therefore influence actions taken, such as re-tweeting, giving a like or replying to the tweet.

The authors of [111] examined the degree to which the age of a user, a feature strongly correlated with readability, influences who people follow on Twitter, and demonstrated that Twitter users have a higher chance to follow people of similar age. Using standard readability measures in text from Twitter, which constrains tweets to be of at most 140 characters in length, is not a trivial task. The lack of structure and shortness of those texts make standard natural language analysis techniques inefficient. With that in mind, we developed TweetRead, a novel readability assessment tool specifically designed for tweets. TweetRead takes advantage of social information, such as hashtags or mentions, for predicting the text complexity levels of tweets. Furthermore, in order to highlight the usefulness of such a tool in social networking environments, we developed a simple, yet effective, hashtag recommendation strategy that takes advantage of TweetRead-generated complexity levels of tweets to inform the hashtag recommendation process.

A.2 TweetRead

TweetRead's goal is to estimate readability of any given tweet T. However, traditional Natural language processing techniques are known not to work properly on short and unstructured text such us the one contained in tweets. Therefore, TweetRead avoids using traditional NLP strategies and relies on simpler models based on content and tweet-specific information. TweetRead is based on a logistic regression technique[1]

[1]We empirically verified that among numerous supervised techniques, logistic regression was the most promising one.

that fuses simple indicators describing T from different perspectives and determines its text complexity. The indicators considered by TweetRead are described below:

- **Extended Flesch.** Flesch reading ease formula) see Figure A.3) is widely used by teachers for estimating readability of texts for their students. However, this formula requires the input text to be sufficiently long to give accurate predictions. Given that tweets are only 140 character long, the accuracy of Flesch is low for predicting the readability of tweets, as shown in our assessment in Figure A.2. To address the issue of the textual content, we explore various strategies that consider tweets that may have similar readability levels. As the number of tweets considered increases, so does the amount of text we have, increasing with it the expected precision of Flesch. We considered 3 tweet groups that may serve as indicators of readability:

  - **Other tweets by the user.** A user is expected to be consistent in his writing in terms of readability. Therefore, to determine the readability of a tweet, it may be useful to take advantage of other tweets written by the same user. For this, the average Flesch of all tweets of the user is used as readability predictor.

  - **Tweets by users mentioned** Homophily stands for the nature of users to relate to similar users. It is a principle that widely manifests in social networks, in terms of aspects such as age, hobbies or profession of users. Following this principle, our hypothesis for this feature is that users with same readability also tend to relate to each other more frequently that random users. Therefore, we consider flesch readability of tweets written by users mentioned in the tweet as indicator of the readability of the tweet.

  - **Tweets that contain same hash-tags** Following the same principle homophily, users we hypothesize that users with similar readability also share the hashtags they usually user. Therefore, we consider the Flesch readability of all tweets containing the same hash-tags that the tweet contains.

- **N-gram models.** Studies [111] demonstrate that users of same age, tend to use similar terminology when writing tweets. Considering that age is a very correlated metric to readability, we take advantage of these writing trends for readability prediction. For doing so, we create one feature $fgram_r$ for each existing readability group r. Each of feature $fgram_r$ is intended to measure the similarity of term distribution between the collection d_r of all the tweets of readability r and the given tweet. For doing so, we take advantage of the well known tfIdf formula, considering each d_r a document containing all the tweets of readability r and D the collection of all tweets.

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D) \tag{A.1}$$

$$tf(t, d) = f(t, d)| \tag{A.2}$$

$$idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \tag{A.3}$$

  The $fgram_r$ similarity of a tweet to the group r is computed as the sum of all tfIdf values of all the terms contained on it.

- **Metatags** Tweets contain more information than just raw text, they also contain Twitter specific information such as hash-tags or mentions and social

information such as emoticons. In order to take advantage of this information we also consider frequencies of this tags as predictors for readability.

$$Flesch = 206.835 - 1.015(\frac{totalwords}{totalsentences}) - 84.6(\frac{totalsyllables}{totalwords})$$

Figure A.1: Flesch reading ease

Unlike traditional readability formulas that tend to map readability levels with school grades, to tailor TweetRead to the Twittersphere, we consider six levels of text complexity following Levinston's [74] adult development stages.

A.3 Hashtag Recommendation

Hashtags are character strings used to represent concepts on Twitter, starting with a # symbol. They are a core Twitter feature and serve classification and search purposes. Their unrestricted nature, however, creates difficulties, including the fact that the same concept can be represented by different hashtags, hindering the search process of a concept [110]. For example, tweets related to the Monaco Formula 1 Grand Prix can be searched using #monacoGP, #monacoF1GP or #monacoF1 retrieving different results. Hashtag recommendation aims at identifying suitable hashtags a user can include in his tweet to reduce the space of tags generated [110] and facilitate the ease with which he and other users can locate the corresponding tweet.

Given that (i) the scope of this paper is to validate the importance of considering a text complexity signal to enhance a recommendation task and (ii) multiple and increasingly complex systems have been developed for hashtag recommendation [59],

we base our study on an existing framework for hashtag recommendation presented in [110]. Given a tweet T, the proposed framework identifies existing hashtags to recommend by following two major steps: (1) generate candidate hashtags by recommending hashtags present in similar tweets, using tf-idf a based similarity and (2) rank hashtags from retrieved candidate tweets using different strategies. The strategies presented in [110] include:

- **Similarity**. Prioritizes hashtags included on tweets that have the closes similarity to T, as estimated using the well-known tf-idf similarity measure.

- **Global popularity**. Prioritizes hashtags based on their respective frequency of occurrence on Twitter.

- **Local popularity**. Prioritizes hashtags based on their frequencies of occurrence among the tweets retrieved in response to T.

We enhance the proposed strategies by taking advantage of TweetRead, as follows:

- **TweetRead**. Prioritizes candidate hashtags that have the same or similar text complexity (estimated using TweetRead) with respect to T.

- **PopularityTweetRead**. Prioritizes hashtags based on their frequencies of occurrence among tweets whose readability level is estimated to match T's.

- **SimilarityTweetRead**. Prioritizes candidate hashtags based on their respective ranking scores computed using Similarity only on tweets whose readability level is estimated to match T's .

Table A.1: Comparison of hash-tag recommendation strategies

	Similarity	GlobalPopularity	LocalPopularity	TweetRead	SimilarityTweetRead	PopularityTweetRead
Mean Reciprocal Rank	0.47	0.19	0.40	0.23	0.52	0.50
First Relevant doc. on avg.	2.14	5.14	2.51	4.39	1.93	2.02

A.4 Initial Assessment

In this section, we discuss an initial evaluation on TweetRead, as well as its applicability for suggesting hashtags.

A.4.1 TweetRead

Given that readability of social content is an unexplored area, benchmark datasets that can be used for evaluation purposes are unavailable. For this reason, we built our own dataset. We initially gathered 172M tweets over an 8-month period using Twitter streaming API. For the purpose of this experiment we assume that the age of people exactly corresponds to their readability level, and that each tweet written by a user will have the same readability level as its author. With that in mind, we followed the framework presented in [111], which examines patterns such as "happy xth birthday", for determining the age of Twitter users. In doing so, we eliminated from our dataset, users (and their corresponding tweets) from whom age could not be determined. Thereafter, we grouped labeled tweets into 6 age groups, which translates into a uniformly distributed dataset of 22k tweets with their corresponding readability levels. We followed a 10-cross-fold validation strategy and measured the accuracy of the predicted readability levels with respect to the ground truth. As shown in Table A.2, TweetRead significantly outperforms the baselines considered for this assessment: Flesch [54] and Spache [99], which are two well-known, traditional readability measures. The reported results demonstrate the need for readability

strategies that examine information beyond standard text analysis, if they are meant to be successfully used in the social networking context.

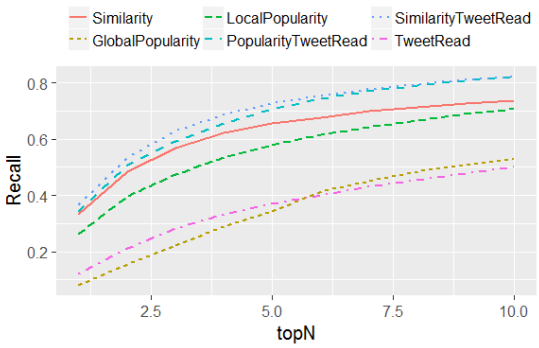
Table A.2: Performance evaluation of TweetRead vs. baselines.

Flesch	Spache	TweetRead
27%	31%	81%

A.4.2 Hashtag recommendation

For evaluating the strategies for hashtag recommendation presented in Section 3, we used the aforementioned dataset. We treated the hashtag of each corresponding tweet as the ground truth. In other words, for each tweet T, we generated the corresponding top-N hashtag recommendations and considered relevant the ones matching the hashtags in T. As in [110], we used the recall measure to evaluate performance and determine to which extend the correct hashtags were recommended within the top N generated suggestions. As shown in Figure A.2, even if readability on its own is not a sufficient factor to suggest hashtags, when combined in-tandem with other content-based and/or popularity strategies, it leads to the improvement of the overall hashtag recommendation process.

Figure A.2: Hashtag recommendation assessment.



To further highlight the improvement achieved by the use of readability, we computed the Mean Reciprocal Rank (MRR) metric for each of the ranking strategy considered. This metric represents in which rank is the first relevant document found in average. We consider hash-tags as documents and the only relevant hash-tag is the one that appears in the ground-truth for the input tweet.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

Figure A.3: Mean Reciprocal Rank (MRR), where Q represents the tested tweets and $rank_i$ the position of the first relevant hash-tag

The results, show that on average the first relevant hash-tag is found at position 1.93 on average for the strategy that combines readability and similarity, being the one that best results achieved in terms of MRR. The best non-readability based strategy is the one that relies on similarity which on average retrieves documents on 2.14 position.

A.5 Conclusion and Future Work

In this paper, we presented TweetRead, a novel readability assessment tool specifically designed to predict the readability of tweets. We also discussed the initial study conducted to demonstrate the benefit of using a readability signal in the hashtag recommendation task, which yielded promising results. In the future, we plan to explore other applications of readability in social networks, such as user recommendation, advertisement targeting or re-tweet prediction. We will also explore techniques to further enhance TweetRead and adapt it to other social networks beyond Twitter.

APPENDIX B

COMPARISON OF FEATURES USED BY A REPRESENTATIVE SAMPLE OF READABILITY ASSESSMENT TOOLS/FORMULAS

	Features	Gonzalez I. (2014) [60]	Denning (2016) [46]	Yaw-Huei Chen (2011) [109]	Amami A. (2008) [16]	Collins-Thompson (2011) [40]	Feng (2009) [51]	Dell'orletta (2011) [45]	Aluisio (2010) [22]	Francois (2012) [56]	Clerq (2016) [44]	Flesch-Kinkaid (1948) [53]	Spaulding (1956) [100]	MultiRead
	Language	EU	EN	CH	AR	CH	EN	IT	EN	FR	NL,EN	Multi	ES	Multi
Shallow	Word Length	Eu	X		X			X	X	X	X	X	X	X
	Sentecen Length	Eu	X		X			X	X	X	X	X	X	X
Morhological	Person Tags													X
	Shortenings	X	?											X
	Modal verbs	X												X
	Case marks	X												X
	Aspect marks	X	X							X				X
	Temporal marks	X	X							X				X
	Mood marks	X	X					X	X	X				X
	Ellipsis marks	X												X
Syntactic	Verb and noun phrases	X							X		X			X
	Subordinates	X					X	X	X		X			X
	Syntactic dependences		X				X	X			X			X
	Linkers	X					X				X			X
	Connectors	X							X		X			X
	POS n-grams													X
	Dependency n-grams													X
	Tree complexity							?			X			X
Semantic	Unigrams (term)			X		X	X	X	X	X				
	Bigrams (term)								X	X				
	Trigrams (term)								X					
	Unigrams (concept)													X
	Synonym Usage													X
	Cohesion									?				X
Other	Book Metadata													X
	Social Data													X
	Web metadata													X

Figure B.1: Features examined by a sample of readability prediction strategies. In this Figure, languages processed by the considered strategies include: Basque (EU), English (EN), Chinese (CH), Arabic (AR), Italian (IT), French (FR), Dutch (NL) and Spanish (ES).