ARASB: Automatic readability assessment system for Basque language

Madrazo I.
Computer Science Department
Boise State University
1910 University Dr
Boise, Idaho
ionmadrazo@boisestate.edu

ABSTRACT

In this paper, we present a novel system for automatic readability assessment for Basque language named as ARASB. ARASB is based on a multinomial bayesian network that makes use of a set of over 4000 features generated for the task. Apart from using traditional features in the area, we present various novel features that describe characteristics of the documents, such as followability of the text or vocabulary enrichment.

In addition, we have analyzed various methods for making good use of the inherent ordinality of the readability assessment task. The proposed system is able to predict 4 different readability levels for the input natural language documents.

Keywords

Readability assessment, Basque, Text tagging

1. INTRODUCTION

Readability assessment is a research area which tries to develop methods for predicting the reading difficulty of a certain text. It has been of great importance in for educational purposes even before automatic natural processing and artificial intelligence had emerged.

At that time manual formulas such as Flesh [Flesch, 1948], Dale-Chall [Chall and Dale, 1995] and Gunning FOG [Albright et al., 1996] showed up as the most used formulas by educators for manually determining text difficulty. Most of dose formulas made use of very simple features such as, average word length or average sentence length.

Once computers and automatic natural language processing started to burst, the area changed drastically its way of developing formulas. More complex features became possible to compute by the help of computers, and supervised learning classifiers started to make appearance on the area.

$$Flesh = 206.835 - (1.015 * AWPS) - (84.6 * ASPW) \quad (1)$$

Figure 1: Flesh's readability formula (AWPS= Average words per sentence; ASPW: average syllabes per word

Even if the main purpose of readability assessment(RA) systems is to serve as help for educator in looking for adequate texts for educator, this is not the only aim for them. RA systems have showed to be useful in areas such as book recommendation [Pera and Ng, 2014] for recommending books adequate for the readers capabilities or in text simplification, to determine whether or not a part of the text needs to be simplified.

This paper is structured as follows. Section 2 provides a description of the related work of the readability assessment area. Section 3 talks about our contribution to the area. Section 4 presents a detailed description of the system. Section 5 presents the results of the experiments done. Finally, sections 6 and 7 present the conclusions and future work of this project.

2. RELATED WORK

In the recent years, different RA systems have been developed with high diversity regarding both languages and features.

For English, [Feng et al., 2010] presented a comparison of the common readability features used for English. [Aluisio et al., 2010] aimed their system for evaluating text simplification methods with a system, that made use of some more elaborated features such as ambiguity in terms. [Feng, 2009] oriented their system for assessing the difficulty level of a text for people with intellectual disabilities, developing some features that were intended to detect how well a text was structured.

For Chinese, [Chen et al., 2011] developed a RA system only based on lexical metrics based on the TF-Idf measure. This metric in conjunction with a mutual information measure was able to determine which terms were most relevant for each of the readability levels. These terms were afterwards used to predict the level of readability for the inputs texts. However, this technique was not topic independent, as once trained for a certain topic the terms were no longer useful for other topics. Previously, [Collins-Thompson and Callan, 2004] developed a system that already tried to solve, the topic dependence problem for Chinese. This system was based on Tf-Idf too and as the authors stated, removing some top scoring words of the Tf-Idf ranking, lead the system to be more independent of the topic. Apparently, the top scoring words were highly specialized words to the topic selected for training

For Arabic, [Al-Ajlan et al., 2008] developed a readability assessment told based on only two features. The features were based on simple ratios based on sentence,terms and letter counts. Those, features were used with a SVM classifier in order to be able to classify text as simple or complex.

For Italian, [Dell'Orletta et al., 2011] presented a readability assessment system aimed for text simplification. Since the text simplification tool the authors were developing was based on sentences. The authors of this system decided,that rather than developing a system for determining text readability, their system would work at sentence level. Therefore, the text simplification tool, would have more information of which sentence needed simplification and which did not. The model generated for sentence level is shown to be generalizable to full text level, by the use of simple averages. The more complex sentences a text have, the more probabilities it have to be complex in overall.

For French,[François and Fairon, 2012] developed a readability assessment system with the foreign language learners in mind. The objective was to determine which features were more important for a foreign language learner to understand a text. In addition, they provided a metric new to the area called adjacent accuracy that tried to measure systems' performance in a more accurate and relevant.

For Basque, for the best of our knowledge, only one system have been developed. Due to the fact that Basque is considered a minority language and shares very little similarities with the most spoken languages. Very little research have been done in the area. Therefore, currently, Errexail [Gonzalez-Dios et al., 2014] is the only system created for Basque readability assessment. This system was aimed for text simplification purposes and was developed to predict two different values, simple or complex. The aim for this was to detect which texts needed some simplification and which texts did not. The system makes use of simple features mostly based on ratios of common Natural language processing tags.

3. CONTRIBUTION

The contribution of this paper is triple:

- We have developed the first readability assessment system for basque that is able to predict more than 2 readability levels.
- We present various novel features to the area of the readability assessment
- We present an analysis of methods that benefit from the inherent ordinality in the class values, for improving performance of the readability assessment systems.

4. PROPOSED SYSTEM

Our proposed system, named as ARASB, is composed of a three step process, a general description of the pipeline can be seen in figure 2.

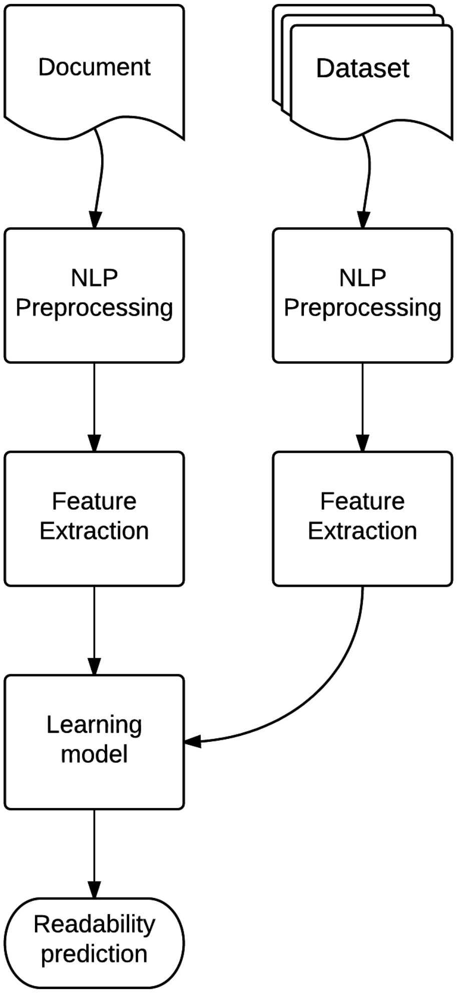

Figure 2: General system pipeline

The first step consists of linguistically analyzing the plain input text. Very little features can be extracted directly from plain text, so the aim of this first step is to preprocess the plain text in order to obtain a enriched text structure.

The second step is the feature extraction phase. This step makes use of previously analyzed text in order to create more elaborated features, over the text.

The last step, is the learning/predicting step. This step makes use of the features generated in the previous step in order to create a model or predict over a previously generated model.

4.1 Text preprocessing

The text preprocessing phase follows a traditional well studied linguistic processing pipeline. The language tags are created in an incremental sequential way, where each analyzer receives an input text already tagged by its predecessor and adds its part of information to it.

This way, in every step, more and more complex tags are generated over the text, creating a an structure that describes the text in various linguistic levels. All the tools used, have been specifically created for Basque language. The pipeline consists of the following analyzers:

1. Morpheus [Alegria et al., 2002] for morpho-syntactic analysis .
2. Eustagger [Aduriz et al., 2003] for lemmatisation and syntactic function identification.
3. [Alegria et al., 2004a] for multiword identification.
4. Ehiera [Alegria et al., 2004b] for Named entities recognition
5. Ixati [Aduriz et al., 2004] for shallow parsing.
6. Maltixa [Bengoetxea and Gojenola, 2010] for sytantactic dependency parsing.
7. MuGak [Aranzabe et al., 2013]for detecting sentence boundaries.
8. [Gonzalez-Dios et al., 2013] for apposition identification.

4.2 Features

In this section we present the features used for the learning model. The features are ordered given their linguistic level. The most elaborated features are presented at the very end.

4.2.1 General features

General features consider the document as whole, and are based on quantities of words and sentences. Four features haves been defined for this group.

- Average number of words per sentence.
- Average number of phrases or syntagms per sentence.
- Average number of letters per word.
- Percentage of terms that appear only once in the text.

4.2.2 Lexical features

Lexical features are features intended to describe the most basic form of a word, also known as the lemma. All the ratios are normalized against the number of words in the text.

- **Ratios of part of speech tags**: verbs, nouns, adverbs...
- **Ratios of person tags**: Who, Who-To_Who, Who-To_Who-Who... (This feature is nonexistent in English)
- **Ratios of Shortenings**: acronyms, abbreviations...
- **Ratios of named entities**: person entities, place entities...
- **Ratios of modal verbs**

4.2.3 Morphological features

Morphological features describe the form changes suffered from lemmas. All the ratios are normalized against the number of words in the text.

- **Ratios of case marks**: inesive, dative, acusative...
- **Ratios of aspect marks**: done or not done marks. (This feature is nonexistent in English)
- **Ratios of time marks**: past, future...
- **Ratios of mode marks**: indicative, subjunctive...
- **Ratios of ellipsis marks**: noun ellipsis, verb ellipsis...

4.2.4 Morphosyntactic features

Five features haves been defined for this group.

- Number of noun phrases per sentence.
- Number of verb phrases per sentence.
- Number of noun phrases per all phrases.
- Number of verb phrases per all phrases.
- Number of appositions per all phrases.

4.2.5 Syntactic features

Syntactic features are features that intend to provide information of the structure of the sentences. All the ratios are normalized against the number of words in the text.

- **Ratios of subordinate sentences**: completive, modal, causal, final...
- **Ratios of syntactic dependences**: subject, object...

4.2.6 Pragmatic features

Pragmatic features provide information regarding the relations between sentences. All the ratios are normalized against the number of words in the text.

- **Ratios of linkers**: causal, final...
- **Ratios of connectors**: conjunction, disjunction...

4.2.7 N-gram model features

N-gram models are a commonly used technique for natural language processing. Those models make use of of probabilities, of a term or more that one term appearing together in a text. Those probabilities are usually used as direct features for learning models.

However, these techniques, fail to achieve topic independence, due to the fact that the features themselves are composed by certain vocabulary, usually inherent to a topic.

The features we present try to follow a similar approach of n-gram models, but instead of using terms our system use probabilities of part of speech tags and syntactic dependencies.

Using this approach we create features such as, a verb followed by a verb, or a noun follows by an adjective. These features, are aimed for providing more accurate information of the structure of the sentences. Our hypothesis states, that the must be very specific structures only existent in high level texts, and some structures only existent in low level texts.

We created this type of features, with all the combinations of part of speech tags and all combinations of syntactic dependencies. Since the syntactic dependencies are modeled as a tree, the probabilities of a certain tag being the parent of another is computed, rather than the probability of a tag being followed by another.

2-gram and 3-gram models have been used for part of speech tags and only 2-grams for syntactic dependencies.

4.2.8 Syntactic tree depth

A syntactic tree is a tree representation of a sentence in which every node represents a phase and the leaf nodes are the words of the sentence. In figure 3 the syntactic tree of the sentence *The chef cooks the soup* can be seen. The example sentence contains a noun phrase and a verb phase, which simultaneously contains another noun phase within it.

Hypothetically, the more complex the syntactic tree of a sentence is, the more complex a sentence is, and therefore, the more difficult it will be to understand.

To model this feature, we implemented a feature that calculates the depth of every tree in a text and and then averages it.

4.2.9 Vocabulary richness

Wordnet [Miller, 1995] is semantically structured database that is commonly used semantic analysis. The database is structured using senses that are connected between each other by sevelar semantic realations, such as synonym, anthonym or hypernym relations. For each sense, its word forms are provided. In figure 4 an example of the wordnet structure can be seen.

The Wordnet database contains around 120,000 different senses, and have been adapted [Agirre et al., 2002] for multiple languages, in which Basque is included.

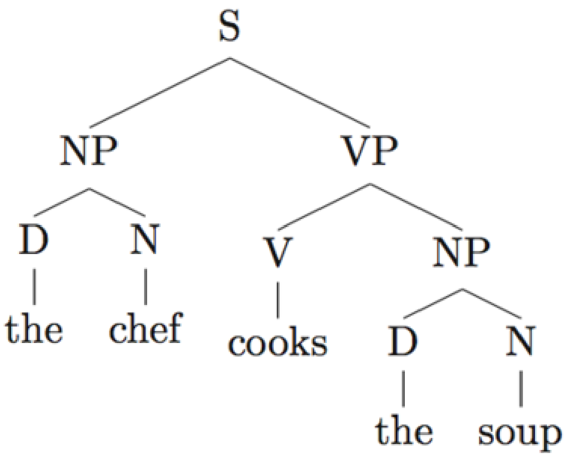

Figure 3: Example of a syntactic tree

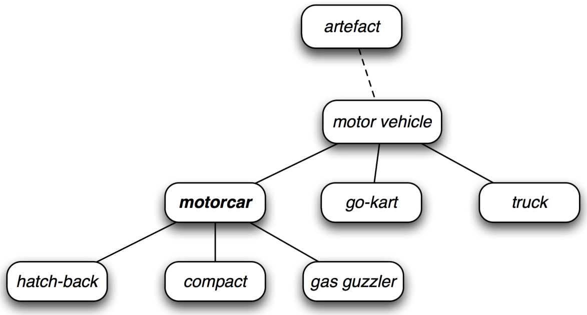

Figure 4: Example of the Wordnet structure

We have used the Wordnet database for developing a feature that describes how semantically rich is a text. Hypothetically, the more rich a text is regarding to vocabulary the more high level it should be.

For computing the measure, the first step is to extract all the senses of the text.Once, this have been done, for each sense, all its word form are retrieved from Wordnet database. Every apparition of each of those word forms is counted and stored.

Finally, in order to calculate how heterogeneous and therefore how rich is the use of vocabulary for each sense, Shannon's entropy [Shannon, 2001] is computed over the probabilities of each wordform for each sense and averaged.

By using this, measure, a sense that has have its appearance in the text by the use of only one term will receive a 0 of vocabulary richness score, and a sense that have appeared in the text in most of its possible forms will get a high vocabulary richness score.

4.2.10 Text followability

The aim of this feature is to measure the ease of followability of the text concepts. For this, the similarity between two sentences will be used. The more similar adjacent sentences are in a text, the softer the changes in concepts are between them and, therefore the easier it is to follow the text ideas. On the other side, the more brusque the changes between sentences are, the more concept changes are between them, and the easier is for the reader to get lost.

For computing this measure, a Latent Semantic Analysis [Deerwester et al., 1990] [Landauer et al., 2013] technique have been used. Latent Semantic Analysis can be used as a similarity measure between two words. LSA, works by creating a matrix of terms and documents, where a 1 is stored in each position where the term appears in that document.

	d1	d2	d3	d4	d5	d6	d7
t1	0	**1**	0	0	0	**1**	0
t2	0	0	0	1	0	0	0
t3	0	1	1	0	0	0	0
t4	0	0	0	0	0	0	0
t5	0	**1**	0	0	0	**1**	0
t6	0	1	0	0	0	0	0
t7	0	0	1	0	0	0	0
t8	0	0	0	1	0	0	1

Table 1: Latent Semantic Anlaysis matrix. t_i stand for diferent terms, and d_i stands for different documents

For calculating the similarity between two term each of their vectors are compared using a cosine similarity. By this approach, having a look at the figure 1, terms t1 and t5 are considered very similar because they co-occur in same documents, while t1 and t2 are considered completely different, because they never co-occur in a same document.

	t1	t2	t3	t4	max
t5	0	0.1	0.5	0	0.5
t6	0	0	0	0	0
t7	0.2	0.4	0.9	0	0.9
t8	0.3	0.4	0.9	1	1
					avg:0.6

Table 2: Computing followability. t1, t2, t3 and t4 represent one sentence and t5, t6, t7 and t8 represent another one. Each cell value represents the LSA score between it row term and its column term.

For computing the followability between two sentences, the procedure in figure 2 is followed. Each term in the first sentence is matched with LSA against each term in the second sentence, calculating the maximum LSA scores for each of the terms in the first sentences. Those maximum values are averaged in order to compute a sentence to sentence LSA similarity. This same procedure is followed for each adjacent sentence pain in the text and averaged.

For training the LSA model, a set of a month of daily articles of a Basque newspaper called *Egunkaria* have been used.

4.3 Learning models

Apart from the commonly used supervised learning algorithms, various new to the area algorithms have been used for developing ARASB. In this section we present the algorithms used, and the main motivation for using them.

4.3.1 CFS Feature Subset Selection

Given that the developed feature set is considerably large(over 4000 features), feature selection has been a important task of the learning task. The algorithm we have used for feature subset selection is the one presented by [Hall and Smith, 1998].

This algorithm makes use of a greedy space search algorithm that uses a correlation based heuristic. The heuristic this algorithm uses, tends to prefer subsets with features with high correlation regarding the class to predict, and with low correlation between each other. This way, this algorithm tries to find a near to optimal subset that reduces the number of features by avoiding redundant information.

4.3.2 Ordinal Classification

The automatic readability assessment task has a inherent ordinarily in the class values that has not usually taken into account for learning purposes. Methods exist that make use of this ordinality in order to improve performance of the classifiers. ARASB makes use of the algorithm presented in [Frank and Hall, 2001]. This algorithm works by creating multiple binary classifiers in a hierarchical way to take profit of the ordinality of the class values.

4.3.3 Cost Sensitive Learning

Cost sensitive learning [Ling and Sheng, 2010], is the set of methods that try to introduce a sense of cost in the learning process. E.g.: For a classifier that tries to predict breast cancer, has not the same cost a false positive, and a false negative. Making a false negative error has a considerably higher cost than a false positive, because the life of a person is in risk in the first case.

A similar thing happens in readability assessment. If a RA system makes an error in predicting the readability level from 1 to 10 for an instance that has a readability value of 1, the cost is not the same if the predicted value is 2 or if the predicted value is 10. The error would be considered larger, when the system predicted a 10.

Therefore, cost sensitive learning have been used in ARASB with the purpose of minimizing the errors magnitude, with the trade-off of sightly lowering the the general accuracy of the system.

5. EXPERIMENT

In this section we present the experiment carried for analyzing the system. Firstly, we present the dataset used for the experiment. Secondly, we show an analysis of the features developed for the system. Finally, a general analysis of ARASB is shown.

5.1 Dataset

Finding a readability leveled dataset for Basque is not a easy task. The education contents for basque are very reduced and most of them are not publicly available.

The dataset we have used for the experiment have been extracted from an online basque learning web page called Ikasbil[1]. This web page contains leveled reading exercises aimed

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

for Basque learners. The leveling is done ussin the European framework of languages, which consists of 6 levels, named as, A1, A2, B1, B2, C1 and C2.

A1	A2	B1	B2	C1	C2
1	24	481	2497	748	39

Table 3: Number of texts per level of readability.

The number of documents for each level can be seen in table 3. The number of text for each level are not homogeneous, and levels A1, A2 and C2 have a very low quantity of texts.

Given that no other dataset was available for the experiment, a decision to discard levels A1 and A2 have been made and level C2 have been complimented with texts from a high level science divulgation web page, called Zientzia [2]. Those text are of enough complexity to be considered of C2 level.

With addition of the documents from Zientzia, a dataset of 1200 documents have been created, with 300 documents for each level of readability, B1, B2, C1 and C2.

5.2 Feature analysis

Two different analysis have been carried for features. Firstly, an overall analysis is presented with the aim of finding the most important features regarding the whole system. Secondly, an analysis of features regarding their importance for each readability level is presented.

5.2.1 *General Analysis*

In figure 4 the top 10 features with highest information gain can be seen. Some of the novel features introduced in this paper seem to have a high importance regarding predicting readability. Features such as n-gram ratios or the more elaborated methods such as vocabulary richness are present in the top 10 of the feature ranking.

In figure 5, it can be seen that when values of the tree depth feature are high (x axis), the presence of high level documents is higher that when its values are low. This shows that the hypothesis we made when developing this feature was correct. Same happens with other developed features.

Information Gain	Feature
0.27807	Ratio of modal verbs
0.20147	Linker + Linker 2-gram ratio
0.20142	Verb + Verb + Verb 3-gram ratio
0.19855	Vocabulary richness
0.19542	Syntactic Tree depth
0.1927	Ratio of verb phases
0.18055	Linker + Noun 2-gram
0.17576	Verb + Verb 2-gram
0.17535	Average length of words
0.16693	Linker(emend) ratio

Table 4: Top 10 features, ordered by information gain. Whole system.

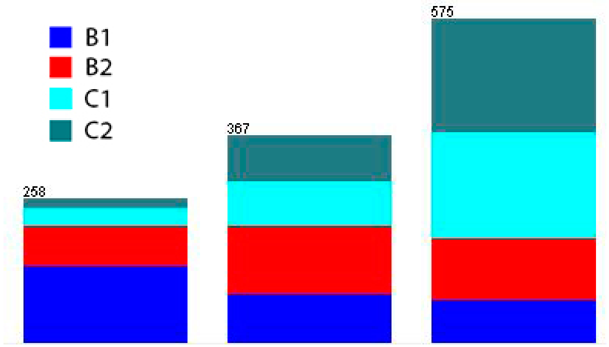

Figure 5: Discretization for syntactic tree depth feature

Group of Features	Avg. Info. gain
Lexical	0.076
Morphologic	0.068
Morfosyntactic	0.097
Syntactic	0.083
Pragmatic	0.093

Table 5: Grouped features ordered by information gain

In the figure 5 the features can be seen grouped by its linguistic level. Morfosysntactic level seems to be the most relevant set of features for determining text readability, followed by pragmatic level.

5.2.2 *Per level Analysis*

The same analysis have been done on a per level basis. For each readability level a binary class have been created that takes as positive example a document form that class and as negative all other documents. Same information gain analysis have been done over those new set, getting diverse results regarding the top 10 features.

For level B1, the top 10 highest information gain features are mostly syntactic dependency based ones. Level B2, seem to be determined by the presence of subordinate sentences, commonly learned at that level. For level C1, the features that have most information gain seem to be related to n-grams, especially related to some certain complex structures such as verb + verb + verb. This level have a light presence of linker features too. Finally, level C2, is strongly influenced by linker related features, 5 out of 10 top features are related to linker presence.

5.3 System analysis

For analysis the system as a whole, the data have been divided into two sets, training (2 out of 3 parts) and test. The training dataset have been used in the algorithm selecting process, using a 10-fold cross validation procedure. Finally, once all the decisions for the system were made, the system have been tested against the test dataset.

[2] http://zientzia.eus/

Class	TP rate	FP rate	Accuracy (Adjacent accuracy)	Recall	F-Measure
B1	0,533	0,140	0,559 (0,853)	0,533	0,546
B2	0,430	0,188	0,433 (0,999)	0,430	0,431
C1	0,750	0,184	0,575 (0,973)	0,750	0,651
C2	0,723	0,009	0,964 (0,978)	0,723	0,827

Table 6: Multinomial naive bayes, detailed results

Algorithm	Accuracy	T-test
Gaussian Bayes net.	57.33 ± 2.02	Base
Multinomial N. bayes net.	61.69 ± 2.48	S.E.
Bayes net. 3 parents.	60.29 ± 2.07	S.E.
C4.5	51.99 ± 2.52	S.W.
1NN	51.81 ± 2.05	S.W.
5NN	56.00 ± 1.02	S.W.
Log. regression	58.90 ± 1.83	S.E.

Table 7: Comparison of learning algorithms. S.E: Statistically Equal: S.W: Statistically Worse. p<0.05

Naive Bayes	Naive Bayes + ordinal classification
60.87 ± 3.83	62.52 ± 4.17

Table 8: Ordinal classification with naive bayes vs. naive bayes

For the test purposes, the feature set that have been used is a subset generated by the CFS feature subset algorithm, which consists of 56 features.

5.3.1 *Selecting learning algorithm*

The system have been tested with different learning models to determine which one fit the best to it. In table 7 a comparison of the tested learning algorithms can be seen. It can be seen that, the three bayesian algorithms and the logistic regression algorithms outperform the algorithms based on nearest neighbor and trees. However, all 4 algorithms are statistically similar in terms of accuracy. Just for the reason of selecting one, for the next experiments, the multinomial naive bayes network algorithm have been chosen.

In table 6, a more detailed information of the performance of the selected algorithm can be seen. Even if the comparison is not fully reliable, the accuracy result shown are at the level of other state of the art systems.

5.3.2 *Ordinal Classification*

In table 8 the results of the comparison between using or not using the ordinal classification previously mentioned can be seen. The use of ordinal algorithms increases the performance in a statistically significant way, validation our hypothesis that such an algorithm would benefit a readability assessment system.

5.3.3 *Cost Sensitive Analysis*

The cost matrix for the cost sensitive algorithm can be seen in table 9. The cost function we used to test our system increases the error cost exponentially, for every unit of distance from the real value.

	B1	B2	C1	C2
B1	0	1	10	1000
B2	1	0	1	10
C1	10	1	0	1
C2	1000	10	1	0

Table 9: Cost matrix

Naive Bayes	Naive Bayes + cost sensitive
60.87 ± 3.83	58.79 ± 4.52

Table 10: Cost sensitive vs. Naive Bayes

The use of the cost sensitive function, decreases the general performance significantly regarding usual accuracy, as it was expected. However, this method increases significantly the adjacent accuracy of the system. Those results can be seen on tables 10 and 11.

5.3.4 *Final test*

After doing different test, the final systems is the composition of CFS feature subset selection, multinomial Naive Bayes, ordinal classification and cost sensitive learning method. The results of the final ARASB system against the test dataset can be seen on table 12. All the results seem to be similar, with a slight decrease, to the ones obtained using the training set with cross validation.

6. CONCLUSIONS

In this paper we presented ARASB, a tool that is able to predict the readability level of a text with a state of the art accuracy. The system is able to predict, over 4 levels of readability, which doubles the quantity of readability levels compared to its predecessor, ErreXail.

The novel features created for the system have shown high prediction ability regarding the class values.

We have introduced two novel approachs to the area of readability assessment, which make use of the inherent ordinality

Naive Bayes	Naive Bayes + cost sensitive
95.07 ± 1.24	97.34 ± 1.33

Table 11: Cost sensitive vs. Naive Bayes (adjacent accuracy)

Class	TP rate	FP rate	Accuracy	Recall	F-Measure
B1	0,463	0,104	0,633	0,463	0,535
B2	0,390	0,226	0,400	0,390	0,395
C1	0,659	0,330	0,435	0,659	0,524
C2	0,583	0,008	0,933	0,583	0,718
Batazbestekoa	0,517	0,186	0,562	0,517	0,523

Table 12: Final system, detailed test results

in the class values with success.

7. FUTURE WORK

As future work, we would like to test our project with a bigger and fuller dataset that contained the missing A1 and A2 readability levels.

We would like to do research in other methods for measuring the followability of the texts, methods based on concept maps that have already been used with success in other areas.

Another interesting experiment would be to take a parallel dataset in two different languages, and do an analysis of the features that affect each language. This way, we could have a better idea of what is important in which language and probably why.

8. REFERENCES

[Aduriz et al., 2003] Aduriz, I., Aldezabal, I., Alegria, I., Arriola, J., Diaz de Ilarraza, A., Ezeiza, N., and Gojenola, K. (2003). Finite state applications for basque. In *Proceedings of EACL'2003 Workshop on Finite-State Methods in Natural Language Processing. Budapest*, pages 13–14.

[Aduriz et al., 2004] Aduriz, I., Aranzabe, M. J., Arriola, J. M., de Ilarraza, A. D., Gojenola, K., Oronoz, M., and Uria, L. (2004). A cascaded syntactic analyser for basque. In *Computational Linguistics and Intelligent Text Processing*, pages 124–134. Springer.

[Agirre et al., 2002] Agirre, E., Ansa, O., Arregi, X., Arriola, J. M., de Ilarraza, A. D., Pociello, E., and Uria, L. (2002). Methodological issues in the building of the basque wordnet: quantitative and qualitative analysis. In *Proceedings of the first International WordNet Conference in Mysore, India*, pages 21–25. Citeseer.

[Al-Ajlan et al., 2008] Al-Ajlan, A. A., Al-Khalifa, H. S., and Al-Salman, A. (2008). 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.

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

[Alegria et al., 2004a] Alegria, I., Ansa, O., Artola, X., Ezeiza, N., Gojenola, K., and Urizar, R. (2004a). Representation and treatment of multiword expressions in basque. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 48–55. Association for Computational Linguistics.

[Alegria et al., 2002] Alegria, I., Aranzabe, M., Ezeiza, A., Ezeiza, N., and Urizar, R. (2002). Robustness and customisation in an analyser/lemmatiser for basque. In *Proceedings of Workshop on" Customizing knowledge in NLP applications". Third International Conference on Language Resources and Evaluation.*

[Alegria et al., 2004b] Alegria, I., Arregi, O., Balza, I., Ezeiza, N., Fernandez, I., and Urizar, R. (2004b). Design and development of a named entity recognizer for an agglutinative language. *IJCNLP-04*.

[Aluisio et al., 2010] Aluisio, S., Specia, L., Gasperin, C., and Scarton, C. (2010). 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.

[Aranzabe et al., 2013] Aranzabe, M. J., De Ilarraza, A. D., and Gonzalez-Dios, I. (2013). Transforming complex sentences using dependency trees for automatic text simplification in basque. *Procesamiento del lenguaje natural*, 50:61–68.

[Bengoetxea and Gojenola, 2010] Bengoetxea, K. and Gojenola, K. (2010). 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.

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

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

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

[Deerwester et al., 1990] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JAsIs*, 41(6):391–407.

[Dell'Orletta et al., 2011] Dell'Orletta, F., Montemagni, S., and Venturi, G. (2011). 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.

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

[Feng et al., 2010] Feng, L., Jansche, M., Huenerfauth, M., and Elhadad, N. (2010). 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.

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

[François and Fairon, 2012] François, T. and Fairon, C. (2012). 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.

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

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

[Gonzalez-Dios et al., 2013] Gonzalez-Dios, I., Aranzabe, M. J., de Ilarraza, A. D., and Soraluze, A. (2013). Detecting apposition for text simplification in basque. In *Computational Linguistics and Intelligent Text Processing*, pages 513–524. Springer.

[Hall and Smith, 1998] Hall, M. A. and Smith, L. A. (1998). Practical feature subset selection for machine learning.

[Irani, 1993] Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning.

[Landauer et al., 2013] Landauer, T. K., McNamara, D. S., Dennis, S., and Kintsch, W. (2013). *Handbook of latent semantic analysis*. Psychology Press.

[Ling and Sheng, 2010] Ling, C. X. and Sheng, V. S. (2010). Cost-sensitive learning. In *Encyclopedia of Machine Learning*, pages 231–235. Springer.

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

[Pera and Ng, 2014] Pera, M. S. and Ng, Y.-K. (2014). 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.

[Shannon, 2001] Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55.