Simple or Complex? Assessing the readability of Basque Texts

Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, Haritz Salaberri
IXA NLP Group
University of the Basque Country (UPV/EHU)
itziar.gonzalezd@ehu.es

Abstract

In this paper we present a readability assessment system for Basque, *ErreXail*, which is going to be the preprocessing module of a Text Simplification system. To that end we compile two corpora, one of simple texts and another one of complex texts. To analyse those texts, we implement global, lexical, morphological, morpho-syntactic, syntactic and pragmatic features based on other languages and specially considered for Basque. We combine these feature types and we train our classifiers. After testing the classifiers, we detect the features that perform best and the most predictive ones.

1 Introduction

Readability assessment is a research line that aims to grade the difficulty or the ease of the texts. It has been a remarkable question in the educational domain during the last century and is of great importance in Natural Language Processing (NLP) during the last decade. Classical readability formulae like Flesh formula (Flesch, 1948), Dale-Chall formula (Chall and Dale, 1995) and The Gunning FOG index (Gunning, 1968) take into account raw and lexical features and frequency counts. NLP techniques, on the other hand, make possible the consideration of more complex features.

Recent research in NLP (Si and Callan, 2001; Petersen and Ostendorf, 2009; Feng, 2009) has demonstrated that classical readability formulae are unreliable. Moreover, those metrics are language specific.

Readability assessment is also used as a preprocess or evaluation in Text Simplification (TS) systems e.g. for English (Feng et al., 2010), Portuguese (Aluísio et al., 2010), Italian (Dell'Orletta et al., 2011), German (Hancke et al., 2012) and Spanish (Štajner and Saggion, 2013). Given a text the aim of these systems is to decide whether a text is complex or not. So, in case of being difficult, the given text should be simplified.

As far as we know no specific metric has been used to calculate the complexity of Basque texts. The only exception we find is a system for the auto-evaluation of essays *Idazlanen Autoebaluaziorako Sistema* (IAS) (Aldabe et al., 2012) which includes metrics similar to those used in readability assessment. IAS analyses Basque texts after several criteria focused on educational correction such as the clause number in a sentence, types of sentences, word types and lemma number among others. It was foreseen to use this tool in the Basque TS system (Aranzabe et al., 2012). The present work means to add to IAS the capacity of evaluating the complexity of texts by means of new linguistic features and criteria.

In this paper we present *ErreXail*, a readability assessment system for Basque, a Pre-Indo-European agglutinative head-final pro-drop language, which displays a rich inflectional morphology and whose orthography is phonemic. *ErreXail* classifies the texts and decides if they should be simplified or not. This work has two objectives: to build a classifier which will be the preprocess of the TS system and to know which are the most predictive features that differ in complex and simple texts. The study of the most predictive features will help in the linguistic analysis of the complex structures of Basque as well.

This paper is organised as follows: In section 2 we offer an overview about this topic. We present the corpora we gathered and its processing in section 3. In section 4 we summarise the linguistic features we

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 334–344, Dublin, Ireland, August 23-29 2014.

implemented and we present the experiments and their results in section 5. The present system, *ErreXail*, is described in section 6 and in section 7 we compare our work with other studies. Finally, we conclude and outline the future work (section 8).

2 Related work

In the last years new methods have been proposed to assess the readability in NLP. For English, Si and Callan (2001) use statistical models, exactly unigram language models, combined with traditional readability features like sentence length and number of syllables per word. Coh-Metrix (Graesser et al., 2004) is a tool that analyses multiple characteristics and levels of language-discourse such us narrativity, word concreteness or noun overlap. In the 3.0 version[1] 108 indices are available. Pitler and Nenkova (2008) use lexical, syntactic, and discourse features emphasising the importance of discourse features as well. Schwarm and Ostendorf (2005) combine features from statistical language models, parse features, and other traditional features using support vector machines.

It is very interesting to take a look at readability systems for other languages as well. Some readability metrics take them into account special characteristics linked to languages. For example, in Chinese the number of strokes is considered (Pang, 2006), in Japanese the different characters (Sato et al., 2008), in German the word formation (vor der Brück et al., 2008), in French the *passé simple* (François and Fairon, 2012) and the orthographic neighbourhood (Gala et al., 2013) and in Swedish vocabulary resources (Sjöholm, 2012; Falkenjack et al., 2013) among many other features. For Portuguese, Coh-metrix has been adapted (Scarton and Aluísio, 2010) and in Arabic language-specific formulae have been used (Al-Ajlan et al., 2008; Daud et al., 2013). Looking at free word order, head final and rich morphology languages, Sinha et al. (2012) propose two new measures for Hindi and for Bangla based on English formulae. Other systems use only machine learning techniques, e.g. for Chinese (Chen et al., 2011).

The systems whose motivation is Text Simplification analyse linguistic features of the text and then they use machine learning techniques to build the classifiers. These systems have been created for English (Feng et al., 2010), Portuguese (Aluísio et al., 2010), Italian (Dell'Orletta et al., 2011) and German (Hancke et al., 2012). We follow the similar methodology for Basque since we share the same aim.

Readability assessment can be focused on different domains such as legal, medical, education and so on. Interesting points about readability are presented in DuBay (2004) and an analysis of the methods and a review of the systems is presented in Benjamin (2012) and Zamanian and Heydari (2012).

3 Corpora

Being our aim to build a model to distinguish simple and complex texts and to know which are the most predictive features based on NLP techniques, we needed to collect the corpora. We gathered texts from the web and compiled two corpora. The first corpus, henceforth *T-comp*, is composed by 200 texts (100 articles and 100 analysis) from the *Elhuyar aldizkaria*[2], a monthly journal about science and technology in Basque. *T-comp* is meant to be the complex corpus. The second corpus, henceforth *T-simp*, is composed by 200 texts from *ZerNola*[3], a website to popularise science among children up to 12 years and the texts we collected are articles. To find texts specially written for children was really challenging. Main statistics about both corpora are presented in Table 1.

Corpus	Docs.	Sentences	Tokens	Verbs	Nouns
T-comp	200	8593	161161	52229	59510
T-simp	200	2363	39565	12203	13447

Table 1: Corpora statistics

Both corpora were analysed at various levels:

1. Morpho-syntactic analysis by *Morpheus* (Alegria et al., 2002)

[1] `http://cohmetrix.memphis.edu/cohmetrixpr/cohmetrix3.html` (accessed January, 2014)
[2] `http://aldizkaria.elhuyar.org/` (accessed January, 2014)
[3] `http://www.zernola.net/` (accessed January, 2014)

2. Lemmatisation and syntactic function identification by *Eustagger* (Aduriz et al., 2003)

3. Multi-words item identification (Alegria et al., 2004a)

4. Named entities recognition and classification by *Eihera* (Alegria et al., 2004b)

5. Shallow parsing by *Ixati* (Aduriz et al., 2004)

6. Sentence and clause boundaries determination by *MuGak* (Aranzabe et al., 2013)

7. Apposition identification (Gonzalez-Dios et al., 2013)

This preprocess is necessary to perform the analysis of the features presented in section 4.

4 Linguistic features

In this section we summarise the linguistic features implemented to analyse the complexity of the texts. We distinguish different groups of features: global, lexical, morphological, morpho-syntactic, syntactic and pragmatic features. There are in total 94 features. Most of the features we present have already been included in systems for other languages but others have been specially considered for Basque.

4.1 Global features

Global features take into account the document as whole and serve to give an overview of the texts. They are presented in Table 2.

Averages
Average of words per sentence
Average of clauses per sentence
Average of letters per word

Table 2: Global features

These features are based on classical readability formulae and in the criteria taken on the simplification study (Gonzalez-Dios, 2011), namely the sentence length and the clause number per sentence. They are also included in IAS (Aldabe et al., 2012).

4.2 Lexical features

Lexical features are based on lemmas. We calculate the ratios of all the POS tags and different kinds of abbreviations and symbols. We concentrate on particular types of substantives and verbs as well. Part of theses ratios are shown in Table 3. In total there are 39 ratios in this group.

Ratios
Unique lemmas / all the lemmas
Each POS / all the words
Proper Nouns / all the nouns
Named entities / all the nouns
Verbal nouns / all the verbs
Modal verbs / all the verbs
Causative verbs / all the verbs
Intransitive verbs with one arg. (*Nor* verbs) / all the verbs
Intransitive verbs with two arg. (*Nor-Nori* verbs) / all the verbs
Transitive verbs with two arg. (*Nor-Nork* verbs) / all the verbs
Transitive verbs with three arg. (*Nor-Nori-Nork*) verbs / all the verbs
Acronyms / all the words
Abbreviations / all the words
Symbols / all the words

Table 3: Lexical features

Among those features, we want to point out the causative verbs and the intransitive or transitive verbs with one, two or three arguments (arg.) as features related to Basque. Causative verbs are verbs with the

suffix *-arazi* and they are usually translated as "to make someone + *verb*", e.g. *edanarazi*, that stands for "to make someone drink". Other factitive verbs are translated without using that paraphrase like *jakinarazi* that means "to notify", lit. "to make know". The transitivity classification is due to the fact that Basque verb agrees with three grammatical cases (ergative *Nork*, absolutive *Nor* and dative *Nori*) and therefore verbs are grouped according to the arguments they take in Basque grammars.

4.3 Morphological features

Morphological features analyse the different ways lemmas can be realised. These features are summarised in Table 4 and there are 24 ratios in total.

Ratios
Each case ending / all the case endings
Each verb aspect / all the verbs
Each verb tense / all the verbs
Each verb mood / all the verbs
Words with ellipsis / all the words
Each type of words with ellipsis / all the words with ellipsis

Table 4: Morphological features

Basque has 18 case endings (absolutive, ergative, inessive, allative, genitive...), that is, 18 different endings can be attached to the end of the noun phrases. For example, if we attach the inessive *-n* to the noun phrase *etxea* "the house", we get *etxean* "at home". The verb features considered the forms obtained with the inflection.

Verb morphology is very rich in Basque as well. The aspect is attached to the part of the verb which contains the lexical information. There are 4 aspects: puntual (aoristic), perfective, imperfective and future aspect. Verb tenses are usually marked in the auxiliary verb and there are four tenses: present, past, irreal and archaic future[4]. The verbal moods are indicative, subjunctive, imperative and potential. The latter is used to express permissibility or possible circumstances.

Due to the typology of Basque, ellipsis[5] is a normal phenomenon and ellipsis can be even found within a word (verbs, nouns, adjective...); for instance, *dioguna* which means "what we say". This kind of ellipsis occurs e.g. in English, Spanish, French and German as well but in these languages it is realised as a sentence; but it is expressed only by a word in Basque.

4.4 Morpho-syntactic features

Morpho-syntactic features are based on the shallow parsing (chunks[6]) and in the apposition detection (appositions). These features are presented in Table 5.

Ratios
Noun phrases (chunks) / all the phrases
Noun phrases (chunks) / all the sentences
Verb phrases / all the phrases
Appositions / all the phrases
Appositions / all the noun phrases (chunks)

Table 5: Morpho-syntactic features

Contrary to the features so far presented, the morpho-syntactic features take into account mainly more than a word. About apposition, there are 2 types in Basque (Gonzalez-Dios et al., 2013) but we consider all the instances together in this work.

[4]The archaic future we also take into account is not used anymore, but it can be found in old texts. Nowadays, the aspect is used to express actions in the future.

[5]Basque is a pro-drop language and it is very normal to omit the subject, the object and the indirect object because they are marked in the verb. We do not treat this kind of ellipsis in the present work.

[6]Chunks are a continuum of elements with a head and syntactic sense that do not overlap (Abney, 1991).

4.5 Syntactic features

Syntactic features consider average of the subordinate clauses and types of subordinate clauses. They are outlined in Table 6 and there are 10 ratios in total. The types of adverbial clauses are temporal, causal, conditional, modal, concessive, consecutive and modal-temporal. The latter is a clause type which expresses manner and simultaneity of the action in reference to the main clause.

Ratios
Subordinate clauses / all the clauses
Relative clauses / subordinate clauses
Completive clauses / subordinate clauses
Adverbial clauses / subordinate clauses
Each type of adverbial clause / subordinate clauses

Table 6: Syntactic features

In this first approach we decided not to use dependency based features like dependency depth or distance from dependent to head because dependency parsing is time consuming and slows down the preprocessing. Moreover, the importance of syntax is under discussion: Petersen and Ostendorf (2009) find that syntax does not have too much influence while Sjöholm (2012) shows that dependencies are not necessary. Pitler and Nenkova (2008) pointed out the importance of syntax. but Dell'Orletta et al. (2011) demonstrate that for document classification reliable results can be found without syntax. Anyway, syntax is necessary for sentence classification.

4.6 Pragmatic features

In our cases, the pragmatic features we examine are the cohesive devices. These features are summed up in Table 7. There are 12 ratios in total.

Ratios
Each type of conjunction / all the conjunctions
Each type of sentence connector / all the sentence connectors

Table 7: Pragmatic features

Conjunction types are additive, adversative and disjuntive. Sentence connector types are additive, adversative, disjuntive, clarificative, causal, consecutive, concessive and modal.

5 Experiments

We performed two experiments, the first one to build a classifier and the second one to know which are the most predictive features. For both tasks we used the WEKA tool (Hall et al., 2009).

In the first experiment we ran 5 classifiers and evaluated their performance. Those classifiers were Random Forest (Breiman, 2001), the J48 decision tree (Quinlan, 1993), K-Nearest Neighbour, IBk (Aha et al., 1991), Naïve Bayes (John and Langley, 1995) and Support Vector Machine with SMO algorithm (Platt, 1998). We used 10 fold cross-validation, similar to what has been done in other studies.

Taking into account all the features presented in section 4, the best results were obtained using SMO. This way, 89.50 % of the instances were correctly classified. The F-measure for complex text was 0.899 %, for simple texts was 0.891 % and the MAE was 0.105 %. The results using all the features are shown in Table 8.

Random Forest	**J48**	**IBk**	**Naïve Bayes**	**SMO**
88.50	84.75	72.00	84.50	**89.50**

Table 8: Classification results using all the features

We classified each feature type on their own as well and the best results were obtained using only lexical features, 90.75 %. The classification results according to their feature group are presented in Table 9. We only present the classifiers with the best results and these are remarked in bold.

Classifier	Random Forest	J48	SMO
Global	74.25	73.50	**74.75**
Lex.	88.00	85.00	**90.75**
Morph.	**82.00**	71.75	75.00
Morpho-synt.	**78.25**	76.25	72.75
Synt.	71.25	**73.75**	67.75
Prag.	67.50	**70.50**	65.75

Table 9: Classification results of each feature type

We also made different combinations of feature types and the accuracy was improved. The best combination group was the one formed by lexical, morphological, morpho-syntactic and syntactic features and they obtain 93.50 % with SMO. Best results are show in Table 10.

Feature Group	Random Forest	SMO
Global+Lex	87.50	**89.50**
Global+Lex+Morph	87.75	**89.00**
Global+Lex+Morph+Morf-sint	89.25	**89.50**
Global+Lex+Morph+Morph-sint+Sintax	87.25	**90.25**
Morph+Morph-sint	**84.25**	82.25
Morph+Morph-sint+Sintax	**83.25**	80.75
Morph+Morof-sint+Sintax+Prag	**83.75**	82.00
Lex+Morph	88.75	**92.75**
Lex+Morph+Morph-sint	**89.25**	**89.25**
Lex+Morph+Morph-sint+Sintax	89.75	**93.50**
Lex+Morph+Morph-sint+Sintax+Prag	88.50	**90.25**
Sintax+Prag	**78.25**	73.50

Table 10: Classification results using different feature combinations

Combining the feature types, SMO is the best classifier in most of the cases but Random Forest outperforms the results when there are no lexical features.

In the second experiment, we analysed which were the most predictive linguistic features in each group. We used Weka's Information Gain (InfoGain AttributeEval) to create the ranking and we ran it for each feature group. In Table 11 we present the 10 most predictive features taking all the features groups into account.

The results of this experiment are interesting for the linguistic studies on Text Simplification. It shows us indeed which phenomena we should work on next. In these experiment we notice as well the relevance of the lexical features and that syntactic features are not so decisive in document classification.

The features with relevance 0 have been analysed as well. Some of them are e.g. the ratio of the inessive among all the case endings, the ratio of the indicative mood among all the verbal moods, the ratio of the adjectives among all the words and the ratio of the ratio of the present tense among all the verbal tenses.

We also performed a classification experiment with the top 10 features and J48 is the best classifier (its best performance as well). These results are presented in Table 12.

To sum up, our best results are obtained using a combination of features (Lex+Morph+Morph-sint+Sintax). We want to remark the importance of lexical features as well, since they alone outperform all the features and 5 of them are among the top ten features.

6 System overview

The readability system for Basque *ErreXail* has a three-stage architecture (Figure 1).

So, given a Basque written text, we follow next steps:

1. The linguistic analysis will be carried out, that is, morpho-syntactic tagging, lemmatisation, syntactic function identification, named entity recognition, shallow parsing, sentence and clause boundaries determination and apposition identification will be performed. We will use the tools presented in section 3.

Feature and group	Relevance
Proper nouns / common nouns ratio (Lex.)	0.2744
Appositions / noun phrases ratio (Morpho-synt.)	0.2529
Appositions / all phrases ratio (Morpho-synt.)	0.2529
Named entities / common nouns ratio (Lex.)	0.2436
Unique lemmas / all the lemmas ratio (Lex.)	0.2394
Acronyms / all the words ratio (Lex.)	0.2376
Causative verbs / all the verbs ratio (Lex.)	0.2099
Modal-temporal clauses / subordinate clauses ratio (Synt.)	0.2056
Destinative case endings / all the case endings ratio (Morph.)	0.1968
Connectors of clarification / all the connectors ratio (Prag.)	0.1957

Table 11: Most predictive features

Random Forest	J48	IBk	Naïve Bayes	SMO
87.75	**88.25**	72.00	83.25	87.00

Table 12: Classification results using the top 10 features

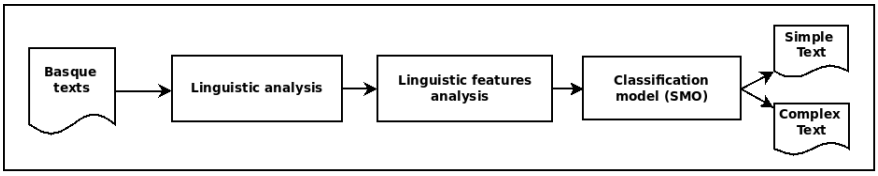

Figure 1: The architecture of system

2. Texts will be analysed according to the features and measures presented in section 4.

3. We will use the SMO Support Vector Machine as classification model, since that was the best classifier in the experiments exposed in section 5. To speed up the process for Text Simplification, we will analyse only the combination of lexical, morphological, morpho-syntactic and syntactic (Lex+Morph+Morph-sint+Sintax) features.

Although the first application of this system will be the preprocessing of texts for the Basque TS system, the system we present in this paper is independent and can be used for any other application. We want to remark that this study, as it is based on other languages, could be applied to any other language as well provided that the text could be analysed similar to us.

7 Discussion

The task of text classification has been carried out by several studies before. Due to our small corpus we were only able to discriminate between complex and simple texts like Dell'Orletta et al. (2011) and Hancke et al. (2012), other studies have classified more complexity levels (Schwarm and Ostendorf, 2005; Aluísio et al., 2010; François and Fairon, 2012). In this section we are going to compare our system with other systems that share our same goal, namely to know which texts should be simplified.

Comparing our experiment with studies that classify two grades and use SMO, Hancke et al. (2012) obtain an accuracy of 89.7 % with a 10 fold cross-validation. These results are very close to ours, although their data compiles 4603 documents and ours 400. According to the feature type, their best type is the morphological, obtaining 85.4 % of accuracy. Combining lexical, language model and morphological features they obtain 89.4 % of accuracy. To analyse their 10 most predictive features, they use Information Gain as well but we do not share any feature in common.

Dell'Orletta et al. (2011) perform three different experiments but only their first experiment is similar to our work. For that classification experiment they use 638 documents and follow a 5 fold cross-validation process of the Euclidian distance between vectors. Taking into account all the features the accuracy of their system is 97.02 %. However, their best performance is 98.12 % when they only use the combination of raw, lexical and morpho-syntactic features.

Aluísio et al. (2010) assess the readability of the texts according to three levels: rudimentary, basic and advanced. In total they compile 592 texts. Using SMO, 10 fold cross-validation and standard classification, they obtain 0.276 MAE taking into account all the features. The F-measure for original texts is 0.913, for natural simplification 0.483 and for strong simplification 0.732. They experiment with feature types as well but they obtain their best results using all the features. Among their highly correlated features they present the incidence of apposition in second place as we do here. We do not have any other feature in common.

Among other readability assessment whose motivation is TS, Feng et al. (2010) use LIBSVM (Chang and Lin, 2001) and Logistic Regression from WEKA and 10 fold cross-validation. They assess the readability of grade texts and obtain as best results 59.63 % with LIBSVM and 57.59 % with Logistic Regression. Since they assess different grades and use other classifiers it is impossible to compare with our results but we find that we share predictive features. They found out that named entity density and and nouns have predictive power as well.

8 Conclusion and perspectives

In this paper we have presented the first readability assessment system for the Basque language. We have implemented 94 ratios based on linguistic features similar to those used in other languages and specially defined for Basque and we have built a classifier which is able to discriminate between difficult and easy texts. We have also determined which are the most predictive features. From our experiments we conclude that using only lexical features or a combination of features types we obtain better results than using all the features. Moreover, we deduce that we do not need to use time consuming resources like dependency parsing or big corpora to obtain good results.

For the future, we could implement new features like word formation or word ordering both based in other languages and in neurolinguistic studies that are being carried out for Basque. Other machine learning techniques can be used, e.g. language models and in the case of getting a bigger corpora or a graded one, we could even try to differentiate more reading levels. We also envisage readability assessment at sentence level in near future.

Acknowledgements

Itziar Gonzalez-Dios's work is funded by a PhD grant from the Basque Government. We thank Lorea Arakistain and Iñaki San Vicente from *Elhuyar Fundazioa* for providing the corpora. We also want to thank Olatz Arregi for her comments. This research was supported by the the Basque Government (IT344-10), and the Spanish Ministry of Science and Innovation, Hibrido Sint project (MICINN, TIN2010-202181).

References

Steven P. Abney. 1991. Parsing by Chunks. In Robert C. Berwick, Steven P. Abney, and Carol Tenny, editors, *Principle-Based Parsing: Computation and Psycholinguistics*. Kluwer Academic.

Itziar Aduriz, Izaskun Aldezabal, Iñaki Alegria, Jose Mari Arriola, Arantza Díaz de Ilarraza, Nerea Ezeiza, and Koldo Gojenola. 2003. Finite State Applications for Basque. In *EACL'2003 Workshop on Finite-State Methods in Natural Language Processing.*, pages 3–11.

Itziar Aduriz, María Jesús Aranzabe, Jose Mari Arriola, Arantza Díaz de Ilarraza, Koldo Gojenola, Maite Oronoz, and Larraitz Uria. 2004. A cascaded syntactic analyser for Basque. *Computational Linguistics and Intelligent Text Processing*, pages 124–134.

David W. Aha, Dennis Kibler, and Marc C. Albert. 1991. Instance-based learning algorithms. *Machine Learning*, 6:37–66.

Amani A Al-Ajlan, Hend S Al-Khalifa, and A Al-Salman. 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.

Itziar Aldabe, Montse Maritxalar, Olatz Perez de Viaspre, and Uria Larraitz. 2012. Automatic Exercise Generation in an Essay Scoring System. In *Proceedings of the 20th International Conference on Computers in Education*, pages 671–673.

Iñaki Alegria, María Jesús Aranzabe, Aitzol Ezeiza, Nerea Ezeiza, and Ruben Urizar. 2002. Robustness and customisation in an analyser/lemmatiser for Basque. In *LREC-2002 Customizing knowledge in NLP applications workshop*, pages 1–6, Las Palmas de Gran Canaria, May.

Iñaki Alegria, Olatz Ansa, Xabier Artola, Nerea Ezeiza, Koldo Gojenola, and Ruben Urizar. 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.

Iñaki Alegria, Olatz Arregi, Irene Balza, Nerea Ezeiza, Izaskun Fernandez, and Ruben Urizar. 2004b. 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.*

Sandra Aluísio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 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.

María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Itziar Gonzalez-Dios. 2012. First Approach to Automatic Text Simplification in Basque. In Luz Rello and Horacio Saggion, editors, *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) workshop (LREC 2012)*, pages 1–8.

María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Itziar Gonzalez-Dios. 2013. Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque. *Procesamiento de Lenguaje Natural*, 50:61–68.

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

Leo Breiman. 2001. Random Forests. *Machine Learning*, 45(1):5–32.

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability Revisited: The New DaleChall Readability Formula*. Brookline Books, Cambridge, MA.

Chih-Chung Chang and Chih-Jen Lin. 2001. Libsvm - a library for support vector machines. The Weka classifier works with version 2.82 of LIBSVM.

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

Nuraihan Mat Daud, Haslina Hassan, and Normaziah Abdul Aziz. 2013. A Corpus-Based Readability Formula for Estimate of Arabic Texts Reading Difficulty. *World Applied Sciences Journal*, 21:168–173.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 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*, SLPAT '11, pages 73–83, Stroudsburg, PA, USA. Association for Computational Linguistics.

William H. DuBay. 2004. The Principles of Readability. *Impact Information*, pages 1–76.

Johan Falkenjack, Katarina Heimann Mühlenbock, and Arne Jönsson. 2013. Features indicating readability in Swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 27–40.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 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.

Lijun Feng. 2009. Automatic Readability Assessment for People with Intellectual Disabilities. *SIGACCESS Access. Comput.*, (93):84–91, January.

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

Thomas François and Cédrick Fairon. 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.

Núria Gala, Thomas François, and Cédrick Fairon. 2013. Towards a French lexicon with difficulty measures: NLP helpnig to bridge the gap between traditional dictionaries and specialized lexicons. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, and M. Tuulik, editors, *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia.*, pages 132–151, Ljubljana/Tallinn. Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Ander Soraluze. 2013. Detecting Apposition for Text Simplification in Basque. In *Computational Linguistics and Intelligent Text Processing*, pages 513–524. Springer.

Itziar Gonzalez-Dios. 2011. Euskarazko egitura sintaktikoen azterketa testuen sinplifikazio automatikorako: Aposizioak, erlatibozko perpausak eta denborazko perpausak. Master's thesis, University of the Basque Country (UPV/EHU).

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36(2):193–202.

Robert Gunning. 1968. *The technique of clear writing*. McGraw-Hill New York.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability Classification for German using lexical, syntactic, and morphological features. In *COLING 2012: Technical Papers*, page 10631080.

George H. John and Pat Langley. 1995. Estimating Continuous Distributions in Bayesian Classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo. Morgan Kaufmann.

Lau Tak Pang. 2006. *Chinese Readability Analysis and its Applications on the Internet*. Ph.D. thesis, The Chinese University of Hong Kong.

Sarah E. Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1):89–106.

Emily Pitler and Ani Nenkova. 2008. 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.

John C. Platt. 1998. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In Bernhard Schlkopf, Christopher J. C Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods-Support Vector Learning*. MIT Press.

Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Satoshi Sato, Suguru Matsuyoshi, and Yohsuke Kondoh. 2008. Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Joseph Maegaard, Benteand Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA).

Carolina Evaristo Scarton and Sandra Maria Aluísio. 2010. Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. *Linguamática*, 2(1):45–61.

Sarah E Schwarm and Mari Ostendorf. 2005. 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.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576. ACM.

Manjira Sinha, Sakshi Sharma, Tirthankar Dasgupta, and Anupam Basu. 2012. New Readability Measures for Bangla and Hindi Texts. In *Proceedings of COLING 2012: Posters*, pages 1141–1150, Mumbai, India, December. The COLING 2012 Organizing Committee.

Johan Sjöholm. 2012. Probability as readability: A new machine learning approach to readability assessment for written Swedish. Master's thesis, Linköping.

Tim vor der Brück, Sven Hartrumpf, and Hermann Helbig. 2008. A readability checker with supervised learning using deep indicators. *Informatica*, 32(4):429–435.

Sanja Štajner and Horacio Saggion. 2013. Readability Indices for Automatic Evaluation of Text Simplification Systems: A Feasibility Study for Spanish. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 374–382, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Mostafa Zamanian and Pooneh Heydari. 2012. Readability of texts: State of the art. *Theory and Practice in Language Studies*, 2(1):43–53.