A Multilingual Application for Automated Essay Scoring

Daniel Castro-Castro[1] , Rocío Lannes-Losada[1], Montse Maritxalar[2], Ianire Niebla[2], Celia Pérez-Marqués[3], Nancy C. Álamo-Suárez[3], Aurora Pons-Porrata[4]

[1]Development Center of Applications, Technologies and Systems, Cuba
{daniel.castro,rocio.lannes}@sc.datys.co.cu
[2]Department of Languages and Information Systems,
University of the Basque Country
{montse.maritxalar,ianire.niebla}@ehu.es
[3]Center for Applied Linguistics, Ministerio de Ciencia, Tecnología y Medio Ambiente, Cuba
{celiap,alamo}@cla.ciges.inf.cu
[4]Center for Pattern Recognition and Data Mining, Universidad de Oriente, Cuba
aurora@cerpamid.co.cu

Abstract. In this paper, we present a text evaluation system for students to improve Basque or Spanish writing skills. The system uses Natural Language Processing techniques to evaluate essays by detecting specific measures. The application uses a client-server architecture and both the interface and the application itself are multilingual. The article also explains how the system can be adapted to evaluate Spanish essays written in Cuban schools.

Keywords: Assessment, Writing, Natural Language Processing in ICALL.

1 Introduction

In recent years, research has been carried out on computer-based Automated Essay Scoring (AES) systems for English ([4, 8, 9, 10, 11, 12]). The AES systems provide students with feedback to improve their writing abilities. Nevertheless, the results so far have been disappointing due to the difficult nature of defining objective criteria for evaluation. Indeed, the evaluation of essays is controversial. In fact, many factors influence the scoring of essays: the topic, time limits, handwriting skills and even the human raters themselves. Most AES systems are based on expert rater evaluations, although some authors [7] use expert writings to develop the evaluation models of such systems.

One of the advantages of AES systems is that they measure all essays using the same scoring model. Moreover, they provide empirical information about the evaluation process itself. In the case of AES systems which use evaluation models based on the criteria of expert raters, the empirical information of the evaluation process provides experts with feedback related to their evaluation criteria. This way, AES systems offer "objective" data to improve on the controversial task of essay evaluation by hand. In this article we address the results obtained from an evaluation of 30 essays and the way these results have influenced the criteria of human raters.

Moreover, we explain the steps followed to define evaluation criteria and how it works in our AES system.

The proposed system uses Natural Language Processing (NLP) techniques to detect specific evaluation measures in the analyzed essays. The application uses a client-server architecture and both the interface as well as the application itself are multilingual. Nowadays, there are few multilingual systems in this field [6]. In this paper, we present a multilingual system, describing the way in which different NLP tools have been integrated for two different languages: Spanish and Basque. Throughout the evaluation, the user gets feedback regarding erroneous linguistic structures, lexical variability and discourse information in both languages as well as information about the grammatical richness of Basque.

In the next section, we will explain the functionality and the architecture of the developed AES system. In section three we define the features detected by the system as well as the NLP tools used. Section four explains the building process of the system's criteria and the evaluation of the system for Spanish essays. Finally, conclusions are outlined in the last section.

2 Functionality and Architecture

We present a bilingual AES system for Spanish and Basque. The system features a Client-Server architecture (see Figure 1).

The server includes a request manager which calls the language modules of the server depending on the language requested by the client. Those language modules are composed of two different types of modules related, respectively, to the linguistic process and to criteria specification. The linguistic process module for **text analysis** and **error management** detects those linguistic features that the client will use to calculate each evaluation measure. This module extracts spelling errors and lexical and discourse information for both languages, as well as syntactic data in the case of Basque essays. The **criteria** specification **module** includes information about the detection of linguistic features (maximum length of a short sentence, word repetition, number of repeated word endings, number of words and different lemmas, specific features of the language such as the written accent in the case of Spanish, etc.). The modules are communicated to the corresponding NLP tools in order to detect the necessary features for each evaluation measure.

We defined three evaluation measures: spelling correction, lexical variability and discourse richness. The linguistic features provided by each evaluation measure are as follows: spelling errors and accentuation (spelling correction), redundancy or word repetition, monotony or repeated word endings, adjective usage (lexical variability), conjugated verbs, sentence length and pronoun usage (discourse richness).

The language modules compute the linguistic features at four different levels: word, sentence, paragraph and text. The result is calculated at one of those levels depending on the linguistic features of the application. For example, word repetition is computed at paragraph level because the client uses the redundancy of the texts at that level. Monotony is also calculated at paragraph level. Pronoun, adjective and conjugated verb usages are considered to be at text level.

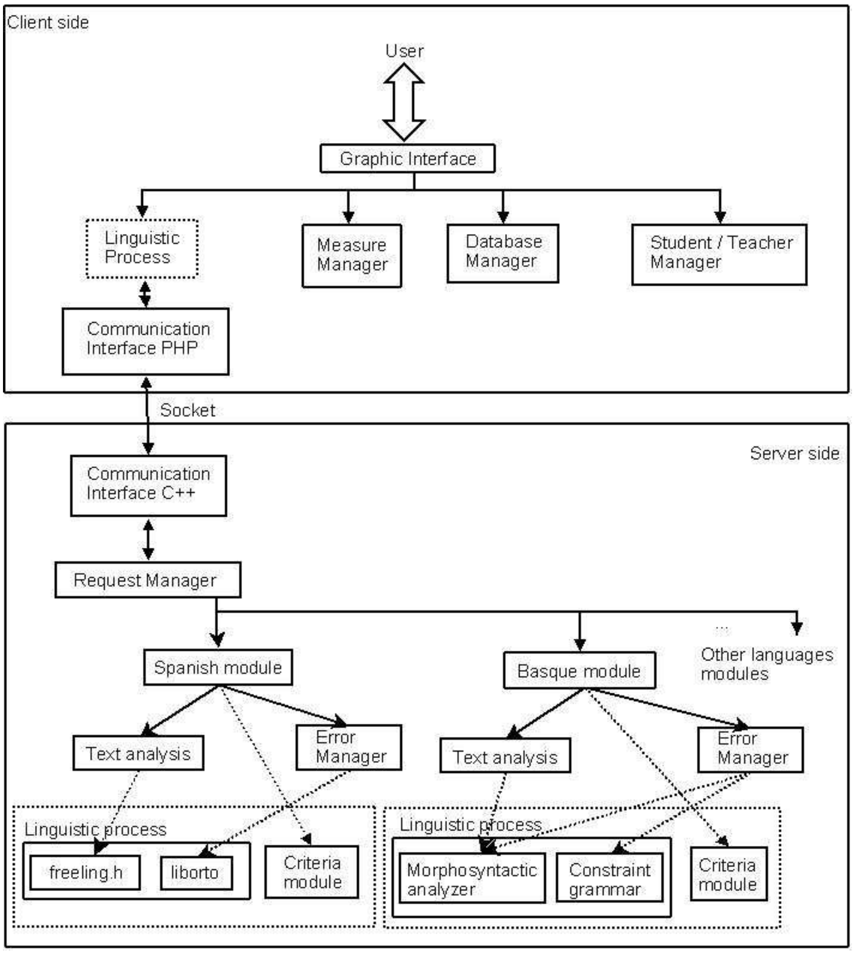


Fig. 1. The architecture.

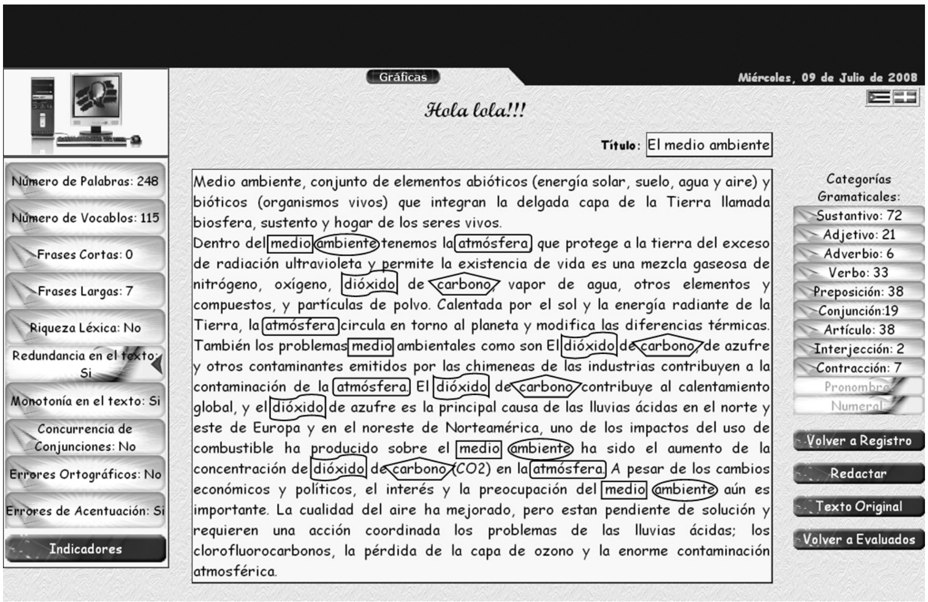
The client interprets the linguistic features calculated by the server in order to compute the evaluation measures that the application will give the user via the results of the evaluation. The application (graphic interface) adapts the interface depending on the language of the essay. The functionalities of the interface are: a text editor to write essays, consultation of previous evaluations and a source of quantitative and qualitative results. Although a score is provided for each evaluation measure, the user can consult the specific linguistic features related to each one. For example (see Figure 2), redundancy (word repetition) in the text is a linguistic feature which influences lexical variability. When the user clicks the button named *Redundancia en el texto (Redundancy in the text)*, the application marks all the repeated words (one color for each different word)[1] in each paragraph. The user employs this kind of

[1] In figure 2 we use geometric figures to represent different colors.

feedback to decide which words are justifiably repeated and which ones must be changed.

There are two types of users: students and teachers. The main difference between them is that teachers can consult the raw formulae used to calculate each evaluation measure. The formulae give information about the linguistic features used when calculating each evaluation measure and the weight given to each feature in the formulae.

The proposed system can be used in both, Windows and Linux operating systems. The client is programmed with *php* while the server modules are written in *C++*.



Título: El medio ambiente

Número de Palabras: 248
Número de Vocablos: 115
Frases Cortas: 0
Frases Largas: 7
Riqueza Léxica: No
Redundancia en el texto: Si
Monotonía en el texto: Si
Concurrencia de Conjunciones: No
Errores Ortográficos: No
Errores de Acentuación: Si

Medio ambiente, conjunto de elementos abióticos (energía solar, suelo, agua y aire) y bióticos (organismos vivos) que integran la delgada capa de la Tierra llamada biosfera, sustento y hogar de los seres vivos.
Dentro del medio ambiente tenemos la atmósfera que protege a la tierra del exceso de radiación ultravioleta y permite la existencia de vida es una mezcla gaseosa de nitrógeno, oxígeno, dióxido de carbono, vapor de agua, otros elementos y compuestos, y partículas de polvo. Calentada por el sol y la energía radiante de la Tierra, la atmósfera circula en torno al planeta y modifica las diferencias térmicas. También los problemas medio ambientales como son El dióxido de carbono de azufre y otros contaminantes emitidos por las chimeneas de las industrias contribuyen a la contaminación de la atmósfera. El dióxido de carbono contribuye al calentamiento global, y el dióxido de azufre es la principal causa de las lluvias ácidas en el norte y este de Europa y en el noreste de Norteamérica, uno de los impactos del uso de combustible ha producido sobre el medio ambiente ha sido el aumento de la concentración de dióxido de carbono (CO2) en la atmósfera. A pesar de los cambios económicos y políticos, el interés y la preocupación del medio ambiente aún es importante. La cualidad del aire ha mejorado, pero estan pendiente de solución y requieren una acción coordinada los problemas de las lluvias ácidas; los clorofluorocarbonos, la pérdida de la capa de ozono y la enorme contaminación atmosférica.

Categorías Gramaticales:
Sustantivo: 72
Adjetivo: 21
Adverbio: 6
Verbo: 33
Preposición: 38
Conjunción:19
Artículo: 38
Interjección: 2
Contracción: 7

Fig. 2. Screenshot of the results provided by the evaluator.

3 NLP Tools and the Detection of Linguistic Features

As seen in figure 1, the system makes use of the open source morphological analyzer, named *freeling* (version 1.5) ([3, 5]), in the Spanish module and our *liborto* library as a basis for the detection of spelling errors in Spanish. In the case of Basque, the morphological analyzer, [1] apart from dividing the words into corresponding lemma and morphemes, provides the morphosyntactic information necessary for the evaluation of the essays. This analyzer is also able to analyze erroneous words. Moreover, the Constraint Grammar formalism [2] includes linguistic rules for the detection of errors at sentence level.

We think that it is worthwhile to provide users with feedback, despite the limitations of NLP tools. For example, in the case of *freeling*, the information related to the part-of-speech is helpful in deciding whether a word must be taken into account when detecting redundancy or monotony. Likewise, the number of adjectives in an essay is another important aspect when evaluating its lexical variability. Hence, the number of adjectives, redundancy and monotony are the three linguistic features used to calculate the lexical variability of the Spanish essays. Discourse richness is based on the number of conjugated verbs and of pronouns, as well as sentence length. We are aware of the limits of this approach since some features, such as the use of conjunctions and coordination or subordination phrases, should also be considered when measuring the discourse level of an essay. However, the *freeling* open source software does not provide this kind of information. As explained in the next section, the raters did not take the mentioned features into account when evaluating the essays. Therefore, they were aware of the limits of the NLP tools before giving a strict evaluation.

4 Evaluation of Spanish Essays

In this section we explain the process followed to develop the system's criteria to evaluate essays written in Spanish in Cuban schools. During the identification of the criteria, the raters changed the formulae used in the evaluation in order to adapt automatic results to their evaluation done by hand. These changes were made in the criteria module. That means that in the future, when we have a wide-coverage analyzer for Spanish, raters will be able to change their formulae in order to take aspects such as subordination and coordination into consideration. The empirical information of the evaluation process provided the raters with specific evaluation criteria feedback and, based on the feedback, the experts changed the weighting assigned to the linguistic features in the formulae.

4.1 The Process

In order to define the criteria module for Spanish, we collected a sample consisting of 30 Spanish essays written by 9th grade students in Cuban schools. In general, the average text length was 237 words.

At the beginning, two experts evaluated a sampling of the compositions in order to define a formula for each evaluation measure. For example, in the case of lexical variability, the experts provided special weights for redundancy and monotony of the text. The lack of adjectives was weighted lower than the previous ones. Indeed, during different interviews with the raters, we realized that we had to give a specific weight to each linguistic feature. That task proved to be difficult as we strived to be as objective as possible.

The raters used a hundred-point scale to evaluate the compositions and a number of points were subtracted each time a linguistic feature was used erroneously. It was not obvious whether the score was the same in the case of each rater, which is related to subjectivity bias. However, by common consent, they defined the number of points to subtract for each linguistic feature. Once they agreed on all linguistic features, we defined the weight that would be given to each. We went on to make up the formulae to be applied in the automatic process. When it came to developing the system, we compared the automatic results and the hand-made evaluations of those 30 essays. We conducted the experiment with three different evaluation measures: spelling correction, lexical variability and discourse richness. In the case of all measures, the totals are counted without considering prepositions, conjunctions or articles.

4.2 The Experiment

For this experiment, spelling correction, lexical variability and discourse richness measures were analyzed. In this case, scores ranged between 1 (the lowest) and 10 (the highest), following the criterion currently used at Cuban schools. The scores provided by our system were compared to scores recorded by hand. The scores reflected precision, recall and F_1.

In this context, we defined these measures as follows:

$$Precision = \frac{Number\ of\ correct\ system\ evaluated\ essays}{Number\ of\ system\ evaluated\ essays}$$

$$Recall = \frac{Number\ of\ correct\ system\ evaluated\ essays}{Number\ of\ manually\ evaluated\ essays}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Table 1 shows the results obtained by the evaluator while factoring in spelling correction. In the table, the first column shows possible test scores. The second column represents the number of texts that raters manually assigned to each score. The third column represents the number of texts for which the evaluator assigned the correct score. The fourth column represents the number of texts to which raters (and not the system) assigned each score. The fifth shows the number of texts to which the system (and not raters) assigned each score. Finally, the last three columns show precision, recall and F_1 values. Likewise, Table 2 shows the results obtained for lexical variability, Table 3 describes the results related to discourse richness and Table 4 shows the results yielded for a global evaluation, where the three mentioned evaluation measures were considered together. The structure of these tables is the same as that of Table 1.

Table 1. Results obtained by the automatic evaluator for spelling correction.

Scores	According to raters	Correctly scored	Missed	Spurious	Precision	Recall	F1
10	26	25	1	2	92.59	96.15	94.33
9	4	2	2	1	75	50	60
Total	30	27	3	3	90	90	90

Table 2. Results obtained by the automatic evaluator for lexical variability.

Scores	According to raters	Correctly scored	Missed	Spurious	Precision	Recall	F1
10	24	22	2	0	100	91.67	95.65
9	6	3	3	2	60	50	54.55
8	0	0	0	2	0	-	-
7	0	0	0	1	0	-	-
Total	30	25	5	5	83	83	83

Table 3. Results obtained by the automatic evaluator for discourse richness.

Scores	According to raters	Correctly scored	Missed	Spurious	Precision	Recall	F1
10	30	30	0	0	100	100	100

Several observations can be made by analyzing the results in Tables 1, 2, 3 and 4. First, despite the difficulty of this task, the system achieves encouraging values of recall, precision and F1 for all evaluation measures. Second, the quality results with respect to spelling correction measure decrease in essays with scores of 9. This is due to the fact that the evaluator recognizes fewer spelling errors than do the raters because it does not identify context errors. In future research, we will improve the *liborto* library to help identify these types of errors.

Table 4. Results obtained by the automatic evaluator for the global evaluation of the texts.

Scores	According to raters	Correctly scored	Missed	Spurious	Precision	Recall	F1
10	13	9	4	1	90	69.23	78.26
9	14	10	4	4	71.43	71.43	71.43
8	2	2	0	3	40	100	57.14
7	1	0	1	0	-	0	-
6	0	0	0	1	0	-	-
Total	30	21	9	9	70	70	70

When it comes to lexical variability, the obtained results are highly dependent on the morphological analyzer. Unfortunately, errors in part-of-speech tagging and unknown words influence these results. Another factor that affects the precision and recall of the evaluator is related to the manual calculation of monotony and redundancy. This is a challenging task for raters, who often detect only minimal repetition.

Another fact that also clearly emerges from the tables is that all essays have high scores, due to the advanced writing ability of the students. We will plan further experiments including a greater number of essays.

5 Conclusions

An essay evaluation system has been presented to help students improve their Basque or Spanish writing skills. This system is the core of the first bilingual web application developed to handle the two aforementioned languages. In addition, it may be easily adapted to other languages thanks to the modularity of the architecture. Moreover, the formulae used for the evaluation can be updated depending on the needs of the human raters. In the near future, we plan on conducting experiments using machine learning techniques to update the formulae.

Analyses of the essays of Spanish students show encouraging results. For evaluation purposes, we have taken three measures into account: spelling correction, lexical variability and discourse richness. Each measure is meant to provide information which must be considered in order to emulate real life scoring as accurately as possible, as a human rater would do. We must recognize that the evaluator is unable to grade in a manner as detailed and elaborated as a teacher would. However, it does provide students with an opportunity to practice their writing skills and it is a way to improve their knowledge of languages, in this case Spanish and Basque.

For future research, we will analyze comparisons with other similar systems, measure the level of rater agreement when evaluating essays and try to include experiments with participants with a wide range of abilities. Coherence and discourse analysis of texts will also be an important line of research in the near future.

Acknowledgments. This research is supported by the Basque Ministry of Education (FOCAD-PRO-2005K2/0001 project).

References

1. Alegria, I., Artola, X., Sarasola, G. K.: Automatic morphological analysis of Basque. In: Literary & Linguistic Computing, vol. 11(4), pp. 193-203. Oxford University Press, Oxford (1996)
2. Aduriz, I., Arriola, J., Artola, X., Díaz de Ilarraza, A., Gojenola, K., Maritxalar, M.: Morphosyntactic disambiguation for Basque based on the Constraint Grammar Formalism. In: Proceedings of Recent Advances in NLP, pp. 282-287. Tzigov Chark, Bulgary (1997).

3. Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., Padró, M.: FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. In: Proceedings of the 5th international conference on Language Resources and Evaluation. Genoa, Italy (2006).

4. Burstein, J., Kukich, K., Wolf, S., Lu, C., Chodorow, M., Braden-Harder, L., Harris, M. D.: Automated Scoring Using A Hybrid Feature Identification Technique. In: Proceedings of the 36th Annual Meeting of the Association of Computational Linguistics, pp. 206-210. Montreal, Quebec, Canada (1998).

5. Carreras, X., Chao, I., Padró, L., Padró, M.: FreeLing: An Open-Source Suite of Language Analyzers. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, vol. 1, pp. 239-242. Lisbon, Portugal (2004).

6. Elliot, S.: IntelliMetric: from here to validity. In: Mark D. Shermis and Jill C. Burstein (Eds.). Automated Essay Scoring: A Cross Disciplinary Perspective, pp. 71-86. Mahwah, NJ: Lawrence Erlbaum Associates (2003).

7. Ishioka, T., Masayuki, K.: Automated Japanese Essay Scoring System based on Articles Written by Experts. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association of Computational Linguistics, pp. 233-240. Sidney, Australia (2006).

8. Kelly, A. P.: General Models for automated essay scoring: Exploring an alternative to the status quo. Journal of Educational Computing Research, vol. 33(1), pp. 101-113 (2005).

9. Landauer, T. K., Laham, D., Foltz, P. W.: Automated Essay Scoring and Annotation of Essays with the Intelligent Essay Assessor. In: Mark D. Shermis and Jill C. Burstein (Eds.). Automated Essay Scoring: A Cross Disciplinary Perspective, pp. 87-112. Mahwah, NJ: Lawrence Erlbaum Associates (2003).

10. Larkey, L.: Automatic Essay Grading Using Text Categorization Techniques. In: Proceedings of the 21st ACM-SIGIR Conference on Research and Development in Information Retrieval, pp.90-95. Melbourne, Australia (1998).

11. Page, E. B., Poggio, J. P., Keith, T. Z.: Computer analysis of student essays: Finding trait differences in the student profile. In: AERA/NCME Symposium on Grading Essays by Computer (1997).

12. Vantage Learning. (n.d.). My Access at http://www.vantagelearning.com.