Bloom's Taxonomy-Based Approach for Assisting Formulation and Automatic Short Answer Grading

Aluizio Haendchen Filho[1,2], Eliane Kormann Tomazzi[2], Rosana Paza[2], Rogério Perego[2], Andre Luis Alice Raabe[1]

[1]Universidade do Vale do Itajaí (UNIVALI)
Mestrado em Computação Aplicada. Itajaí – SC – Brazil

[2]Centro Universitário de Brusque (UNIFEBE)
Núcleo de Inteligência Artificial e Sistemas Inteligentes. Brusque – SC – Brazil

{aluizio.h.filho, rgperego}@gmail.com, ropaza@hotmail.com, raabe@univali.br

Abstract. *This paper presents an approach to enhance automatic short answer grading accuracy by using Bloom's Taxonomy as a reference for questions formulation. We sought to address the semantic aspects related to the answer by using WordNet and Latent Semantic Analysis models, which supported automatic short answer grading with size ranging from a single sentence to a short paragraph. The responses for three questions answered by high school students were graded automatically resulting in a high correlation with teacher grading (0.82, 0.91, 0.80). Another discovery is that automatic correction might vary according to the type of question, the application context and that the representativeness and concision of the expected response.*

1 Introduction

Research on automatic grading has become more relevant with the emergence of virtual learning environments. The computerized correction has numerous benefits. In addition to its low cost, it permits instant feedback and eliminates the work of manual corrections, allowing to assist an expressive number of students. Moreover, it qualifies the writing process, since teacher will have better use of its pedagogical time.

During the school years, the student naturally passes through an evaluation process, which is continuous, cumulative and systematic. Even in the most modern pedagogical conceptions, the application of tests with questions of the discursive type has great relevance, since they evaluate the capacity of reading, interpretation, and writing. However, manual grading of many questions from the considerable number of students demands lots of time from teachers. This lead to a tendency to elaborate tests with only multiple-choice questions, which present limitations in the development of the teaching and learning processes. In multiple-choice questions formats, the answer is standardized and equal for all, that is, a proposition or a question presented is either true or false, and there is no question possibility. According to Kleinke (2008), multiple-choice tests make the student seek only the right alternative, and assessments with discursive questions

DOI: 10.5753/cbie.sbie.2018.238

allow the candidate to develop reasoning and critical thinking about answers.

In the discursive question, the student is free to direct the answer and must construct and formulate it in an adequate language. Just as in essay writing, all kinds of discursive questions evaluate the candidate's ability to interpret and produce texts. Short answer questions focus on content, while essays focus on style [Burrows, Gurevych, and Stein 2015].

This evaluative configuration also promotes greater use of the pedagogical time by the teacher, who can elaborate more dynamic lessons for knowledge constructing. This would allow the teacher to act more effectively in the students' individual learning, amplifying the process of writing and reading. These elements had historically low proficiency in most Brazilian schools.

This educational scenario, according to data released by the National Institute of Educational Studies and Research (INEP), has indicated that the reading and writing skills of students are stagnated in the country between the years 2014 and 2016. More than half of students in the 3rd year of elementary school have an insufficient level of reading, that is, difficulty in interpreting a text and mathematics. In the case of writing, 33.9% of the students show insufficient proficiency. The data show that the policies developed so far have not produced effective results. The same situation of insufficiency of results observed in 2014 is repeated in 2016. According to the minister of Education, there has not been evolution or significant improvement [Peduzzi 2017].

Research on automatic grading of discursive questions and essays is being developed since the 1960s, but only in the current decade, they are achieving the accuracy needed for practical use in educational settings. For end users to have confidence, the challenge is to develop robust and accurate assessment systems close to human evaluators (Santos, 2016). In the literature, there are several studies on the automatic grading of discursive questions [Burrows; Gurevych; Stein 2015].

Given the large number of published techniques but with application in specific contexts, new studies are required comparing existing techniques [Burrows & Stein 2015; Ziai & Meurers 2012]. In some studies, new parameters are tested in order to refine existing models. In [Santos 2016], the standard LSA technique was improved and an accuracy of 84.94% was obtained in a corpus of 349 responses, similar to the 84.93% agreement among the human evaluators. Mohler & Mihalcea (2009) carried out a similar research, where they explored knowledge and corpus-based techniques on 21 questions and 637 answers written by Computer Science students. The best results were obtained using LSA with a corpus of Wikipedia articles pertinent to a specific domain and a refinement based on the best answers ($r = 0.5099$).

One of the identified gaps is the absence of papers that evaluate the impact of the question formulation methodology and the reference answer on the performance of existing approaches. In order to fill this gap, we propose an approach based on the Bloom's Taxonomy for formulation and automatic answer grading, aiming to improve the accuracy of the educational systems in the evaluation of learning. It also evaluated the most effective format of the expected responses for automatic short answer grading. An application was developed for this purpose, and used in a practical experience held in a

high school environment.

2 Background

The main basis of our research for the formulation of discursive questions is the Bloom Taxonomy [Bloom et al. 1973]. Bloom's Taxonomy establishes a hierarchical classification of the learning stages and defines the objectives to be achieved at every stage. The cognitive domain was classified into six categories: knowledge, understanding, application, analysis, synthesis and evaluation. Figure 1 [Duquesne University 2017] shows the main verbs belonging to each category. The action verbs below each category illustrate the kinds of activities a test item can evaluate.

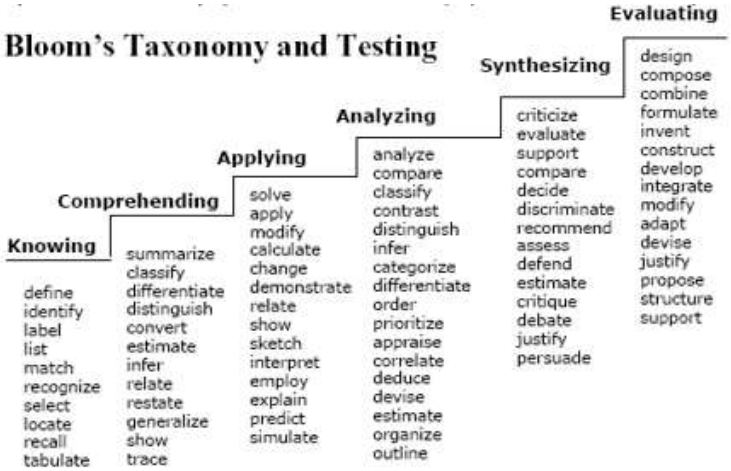

Figure 1 - View of cognitive categories and Bloom's Taxonomy verbs

According to Santana Júnior (2008), the Bloom's categories represent cumulative intellectual processes, which follow a hierarchical line, requiring the individual to master an earlier goal before reaching the next. Each step provides support for accessing to a higher level of knowledge, showing a dependency relation among them. The use of verbs in question construction can make it easier for students to understand what is expected.

In their research, Athanassiou, McNett, and Harvey (2003) state that taxonomy was used to help students to appreciate the conceptual richness of the study material. It acts as a self-assessment tool, also working to help students to understand the conceptual complexity which a given study may present. The authors further argue that this construct is the basis for an educational model based on competence.

When hierarchizing educational behaviors from the simplest to the most complex, Bloom et al. (1973) argue that this idea is based on the conception that a simple behavior integrated with equally simple ones, becomes a more complex one. Unlike the multiple-choose test items, which are ideal for testing the broad knowledge of content in a relatively short period of time, the discursive questions are best suited for assessing the level of learning. By nature, they require more time for students to think, organize, and compose their answers.

This gap in the reading and writing aspects, at the beginning of the school process, indicates that an educational policy designed to improve these indices have failed. This

gap is carried on to High School and University students. This imposes a great challenge to elementary schools and teachers. In this educational scenario, this situation is aggravated by premature evasion, revealing a fragile educational system.

Therefore, educational policies must be rethought, as well as the evaluation process. The teacher, while thinking about assessment regarding the exams and tests, should consider the deficit of writing and reading, not as a stated fact, but as a possibility of transformation. The production of discursive questions can aid to promote this change, whether committed to student learning, as considered in Bloom's Taxonomy. Next, is partially presented an application with all the steps for the construction of this educational proposal, from the elaboration to correction.

3 A Bloom Taxonomy-based application for question creation and grading

For the formulation of questions, creation, and correction of the tests based on the Bloom's Taxonomy, an application was developed. The creation of tests includes the following steps: (i) choice of a subject or component of the curriculum, which may be high school or higher education; (ii) insert a text or figure to contextualize the question, if necessary; (iii) choosing the category of the cognitive domain of taxonomy and the verb related to this category; (iv) insert the statement of the question; (v) inform the expected response and keywords; (vi) review of questions and expected responses; (vii) selection of the questions to compose the test. Figure 2 presents a partial view of the graphical interface for question creation.

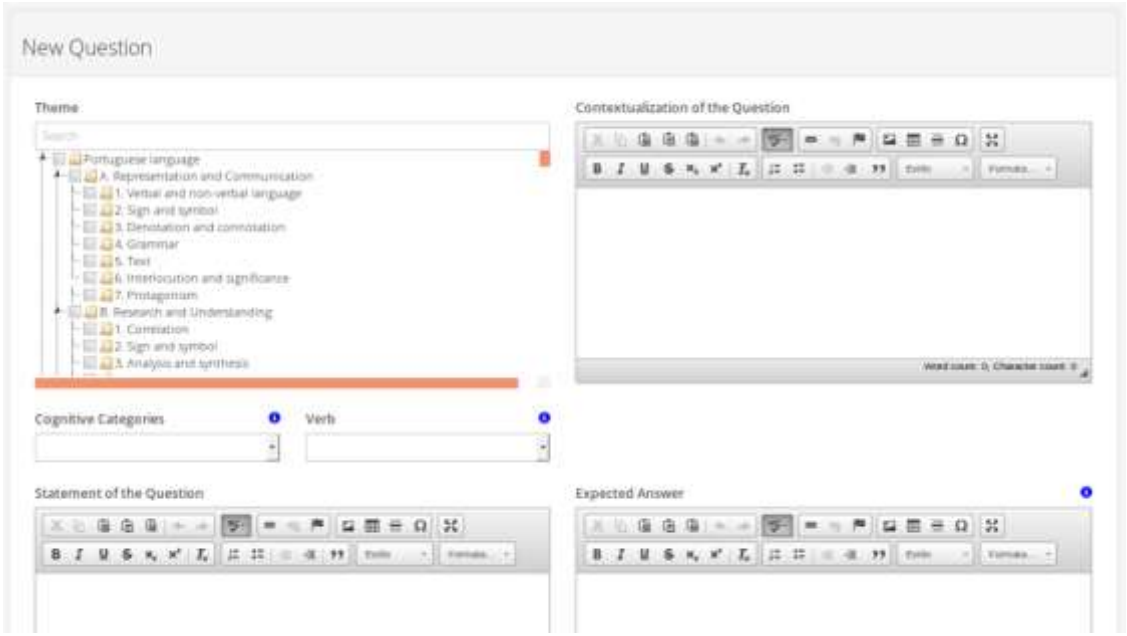

Figure 2 - Partial view of GUI for question creation

The example presents the curricular components of secondary education in Brazil, which are inserted in a hierarchy with the four Knowledge Areas defined by the National Curricular Common Base [BNCC 2015]: Languages, Mathematics, Human and Nature Sciences. These areas, defined in the CNE / CEB Ordinance No. 11/201027, "favor the

communication between the knowledge of the different curricular components". After selecting the subject, it is possible to insert a contextualization text or image for the question, as can be seen in the window on the right side of the figure.

The next step in formulating the question is the choice of the cognitive domain category of Bloom's Taxonomy. When selecting a category, the system displays the commands or verbs that can be used in the elaboration of the statement. After selecting the main verb or the command of the question, it is possible to insert the statement, which must contain the selected verb. Each statement should contain only one command; this procedure facilitates the correction since each command can have specific punctuation value. The system checks to validate the statement. After entering the statement, the question formulator should enter the expected response and relevant keywords for the response. In the application, a graphical interface provides an input box for insertion. The expected response can be reported in text format or in topic format.

The application provides graphical interfaces and resources to assemble tests, streamline and facilitate the correction of questions by human reviewers. It also provides resources to send the correction directly to students with feedback. The goal is to reduce the correction time and eliminate spent and paper handling.

4 Case Study

We present in this topic a case study in which the application was used to formulate questions, create, apply and correct exams. In the first stage of the project, the experiments were done to validate the solution with short answers, with a maximum of 300 characters.

4.1 Formulation of questions and expected answer

For the elaboration of the questions, a training seminar for teachers at the Amplo High School in Brusque/SC Brazil was ministered, dealing with good practices for its formulation. At the seminar, the following topics were addressed: (i) how to contextualize the question; (ii) how to formulate the questions using Bloom's Taxonomy; and (iii) how to inform the expected answer to each question.

In the seminar, there were 12 high school teachers of different areas of knowledge, including History, Geography, Portuguese, Physics, Chemistry, Biology, Sociology, and Philosophy. Each teacher inserted 3 to 5 questions in their area of knowledge regarding the content seen in class in the 1st year.

A database was created with 45 questions, which were reviewed by researchers in pedagogy based on ENEM / ENADE criteria. Among the validated questions, 21 were selected to compose a multidisciplinary test with a time between one and two hours. Twenty seven (27) students answered the test from 1st to 3rd year, totaling a set of 477 valid answers and 90 blanks or with an answer not directed to the question.

For the purposes of this study, three questions were selected: one from the first, and two from the third cognitive category of Bloom's taxonomy. The choice of a simpler and two slightly more complex cognitive categories allowed us to obtain parameters to

analyze the results of automatic correction in different cognitive categories. Some questions of the early cognitive categories were chosen because one of the objectives of the study is to measure the accuracy for very short, short and medium questions.

For these questions, 27, 24 and 25 valid answers were collected, totaling 76 answers, and the average number of words per response was 30, 11 and 30, respectively. Each answer was given a score of 0 to 10 by two teachers, who are experts in the field. The procedures for inclusion of the questions, assembly, application of the test, and correction of the answers were described in Section 3. Table 1 presents teacher reference questions and answers to these questions.

Table 1. Questions used in the research corpus

Geography	**Question 1** **Statement**: Explain the role of DST (daylight saving time) in Brazil. **Reference answer**: Save energy by making better use of light, which extends for longer especially in southern Brazil. This region is more towards the sun due to the axis of inclination of the Earth.
Portuguese Read the text below: Marcos, 31, was arrested on Sunday afternoon, in São Paulo city, due to an active arrest warrant from the Paraná Court of Justice. The officers of the Tactical Patrol Squad were patrolling the streets when they saw Marcos riding a motorcycle. When consulting the system, they verified the presence of an arrest warrant and that the vehicle was in an irregular situation, which was taken to a detention garage. Marcos was also taken to Advanced Prison Unit (UPA) of São Paulo during the afternoon.	**Question 2** **Statement:** Identify the verbal voice that prevails in the text. **Reference answer:** Passive voice. **Question 3** **Statement:** Justify the use of the predominant verbal voice. **Reference answer:** Passive voice; highlight of the occurring action; generalization of the subject.

4.2 Automatic short answer grading

The data obtained in this case study were used in an automatic short answer grading experimentation. The approach comprises three steps: (i) collecting and storing the information, as presented in previous topics and subtopics; (ii) processing, described in Passero et al. (2016); and (iii) analysis of the results, summarized below.

Before commenting on the analysis of the results, some preliminary processing information is needed to understand the synthesis of analysis. In the processing phase, LSA (Latent Semantic Analysis) models [Burrows, Gurevych and Stein 2015] were used; [Landauer and Dutnais 1997] with the preprocessed Wikipedia base using the open library Semantic Vectors [Widdows and Ferraro 2008], with the dimensions [200, 250, 300, 350, 400, 450, 500] and the 441,000 most frequent terms (Frequency> = 10). The model with 350 dimensions was selected to represent the LSA since it presented the best results.

In addition to the LSA, the WordNet model was used, whose algorithms used to calculate the similarity were adapted from the free WordNet Similarity library [Pedersen and Michelizzi 2004, Fellbaum 1998]. The Apache Jena framework was used to load OWN-PT data into main memory using SPARQL (SPARQL Protocol and RDF Query Language). The algorithms used to calculate similarity were adapted from the free WordNet Similarity library [Pedersen & Michelizzi 2004].

In the result analysis step, the objective was to check the semantic similarity

between the student's answer and the teacher's reference answer. Passero et al. (2017) showed that different similarity models might perform better on different types of questions. In their study, measures based on WordNet showed better results in a shorter reference response; on the other hand, the LSA model was more efficient in longer responses. The agreement among the evaluators was calculated using the Pearson Correlation (r), MAE (mean absolute error) and RMSE (root mean squared error) measures. Table 2 shows the values obtained.

Table 2. Agreement among the evaluators

Question	Correlation (*r*)	MAE	RMSE
1	0,82	1,792	2,953
2	0,84	1,440	2,300
3	0,71	2,370	3,040
Total	0,70	1,882	2,840

It was also observed that human reviewers provided the same score in 27 responses (35.53%), differentiated in one or two points in 25 (32.89%), three to five points in 21 (27.65%) and six to ten points in 3 (3.95%). Table 3 presents the results of correlation of teacher grading with LSA and WordNet models.

Table 3. Summary of the results obtained with the methods used

	Question								
	1			2			3		
Model	*r*	MAE	RMSE	*r*	MAE	RMSE	*r*	MAE	RMSE
LSA	0,829	1,296	1,599	0,450	3,480	4,015	0,757	1,083	1,581
WordNet	0,713	1,667	2,010	0,910	1,600	1,980	0,804	1,125	1,458

The presented summary indicates that the performance indexes found are similar to the agreement between evaluators. The isolated LSA model obtained a better result in Question 1, which had a longer explanatory reference response (32 words). For this question, the LSA model presented the highest correlation (0.829). The WordNet had better results in Questions 2 and 3. Question 2 had the lowest reference response, requiring the student to mention the expression "passive voice" in his answer, the method had the best correlation (0.910). For question 3 (11 words), the method reached a correlation of 0.804.

6 Discussion of results

The presented approach has shown to be feasible to support automatic short answer grading with size ranging from a single sentence to a short paragraph. In practice, it was observed that the obtained results in the automatic correction might vary according to the type of question, the application context and that the representativeness and concision of the reference response can affect the performance of the evaluated techniques.

In the manual grading of the students' answers, carried out by two human evaluators, several orthographic errors were identified that, despite indicating a lack of

proficiency in the Portuguese language, did not prevent the evaluator from recognizing the presence of the expected concepts and considering the answer as correct. An automatic corrector to avoid harming automatic grading adjusted the identified spelling errors. In preliminary tests, it was found that the similarity analysis methods used returned a lower or zero index when the response words were spelled incorrectly.

Regarding the formulation of the questions, it was verified that:

a) Bloom's Taxonomy facilitated the elaboration of understandable discursive questions and properly related to the educational objectives of the curriculum.
b) In the formulation of questions, the teachers stated that Bloom's Taxonomy was useful to clearly define the resulting learning intended to be assessed by the question.

Related to the expected answer, it was found that:

a) Establishing the expected answer, identifying its key parts to be used as a reference in the correction, is critical to increasing the accuracy of automatic grading.
b) The verification that there may be different alternatives of answers considered correct focuses on the common points of all the answers, and the consequent elimination of unnecessary words.

The specialist performed tests with the answers provided by the teachers, prior to the validation and adjustment and a reduction of up to 12% in accuracy was verified for the WordNet model when the expected answer contained unnecessary words. The WordNet-based metrics require that all expected concepts be addressed in the student's response to a 100% index. For the LSA model, the unadjusted answers also resulted in reduced accuracy and correlation (-2% and -0.13). For example, adjusting the answer "The predominant voice is the passive voice" to "Passive voice", students who do not include a word related to the concept "predominate" would not be harmed.

The usability of the application have not created obstacles for teachers since those who used it for the first time expressed a desire to use it regularly. Students also gave a very positive feedback, because the test in digital means allows the answers to be adjusted and improved without erasures. In future applications, a spell checker could be enabled in the graphical interface, used by the student without impairing to the evaluation, except in cases where the writing domain is important.

7 Conclusion and future work

In this paper, we present an application based on Bloom Taxonomy to facilitate the formulation process and automatic short answer grading. The approach sought to address the semantic aspects related to the answer and presented relevant results in our case study, with questions that use the lower cognitive categories of the Bloom Taxonomy. Using the taxonomy to produce statements and answering patterns is in line with Bloom's goals of creating a common vocabulary for expressing cognitive qualities.

The main contributions of this work are: (i) the adoption of Bloom's taxonomy in the context of discursive issue corrections, contributing to improvements in the process

of question formulation; (ii) improvements in accuracy of the automatic grading system is achieved when the expected answers are well formulated; and (iii) present an approach that demonstrates effectiveness for use in the process of automatic short answer grading.

Following the first stage of the work, our efforts will be directed towards the correction of longer answers involving the higher cognitive dimensions of the Bloom Taxonomy. For automatic grading of the higher cognitive categories, it is necessary to consider also the syntactic and pragmatic aspects of the answer. For this, we will use the following resources:

a) Enhance the semantic understanding of the text by relating the Bloom Taxonomy verbs to packages of words that fit the idea involved in the verb or command of the question.
b) Relate Bloom's Taxonomy verbs with discourse markers and check the existence of these markers in the answer. The connectives, which encompass linguistic elements belonging to different classes of words (conjunctions, adverbs or interjections), are part of discourse markers.
c) Consider the understanding of the discourse, in addition to the semantic understanding, using techniques for cohesion analysis and textual coherence, used in our work of automatic grading of essays [Haendchen Filho et al. 2018].

References

Andrade, D.; Campos, M. (2010) Análise do processo cognitivo na construção das figuras de Lissajous. Revista Brasileira de Ensino de Física, v. 27, n. 4, p. 587-591, 2010.

Athanassiou, N.; McNett, J. M.; Harvey, C. (2003). Critical Thinking in the Management Classroom: Bloom's Taxonomy as a Learning Tool. Journal of Management Education, 27(5), 533-555.

Bloom, B. S. et al. (1973). Taxionomia de objetivos educacionais. Porto Alegre: Globo.

BNCC. (2017). Base Nacional Comum Curricular. Ministério da Educação, 2017.

Burrows, S., Gurevych, I. e Stein, B. (2015). The eras and trends of automatic short answer grading. International Journal of Artificial Intelligence in Education, v. 25.

Duquesne University (2017). Strengths and Dangers of Essay Questions for Exams. Disponível em:http://www.duq.edu/about/centers-and-institutes/center-for-teaching-excellence/teaching-and-learning/strengths-and-dangers-of-essay-questions-for-exam

Fellbaum, C. (1998). WordNet: An electronic lexical database. Language, Speech, and Communication. MIT Press, Cambridge, USA.

Ferraz, A. P. C. e Belhot, R. V. (2010). Taxionomia de Bloom: revisão teórica e apresentação das adequações do instrumento para definição de objetivos instrucionais. Gestão e Produção, São Carlos, v.17, n.2, p. 421-431.

Haendchen Filho, A., Prado, H. A., Ferneda, Edilson, Nau, J. (2018) An approach to evaluate adherence to the theme and the argumentative structure of essays. In: Knowledge-Based and Intelligent Information & Engineering Systems, Proceedings of the 22nd International Conference, KES-2018, Belgrade, Serbia. London: Elsevier

B.V., 2018. v. 126. p. 788-797.

Kleinke, M.U. (2008). Tipos de prova (objetivas e discursivas): a interdisciplinaridade como elemento articulador. XXXII SAESUNN – Seminário de Acesso ao Ensino Superior das Universidades do Norte e Nordeste, 2008.

Landauer, T. K. e Dutnais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. Psychological Review, 104(2), 211–240.

Mohler, M. & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. In EACL '09 Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Edited by Association for Computational Linguistics Stroudsburg, PA, USA, Pages 567-575.

Passero, G., Haendchen Filho, A. & Dazzi,R. (2016). Avaliação do Uso de Métodos Baseados em LSA e WordNet para Correção de Questões Discursivas. Proceedings of the XXVII Brazilian Symposium on Computers in Education (SBIE 2016. (pp. 1136–1145)

Pedersen, T., Patwardhan, S. e Michelizzi, J. (2004). WordNet::Similarity: Measuring the Relatedness of Concepts. In Demonstration Papers at HLT-NAACL 2004. HLT-NAACL- -Demonstrations '04. Association for Computational Linguistics.

Peduzzi, Pedro. Mais de 50% dos alunos do 3º ano tem nível insuficiente em leitura e matemática. Repórter da Agência Brasil. Brasília. 2017. Disponível em: <http://agenciabrasil.ebc.com.br/educacao/noticia/2017-10/mais-de-50-dos-alunos-do-3o-ano-tem-nivel-insuficiente-em-leitura-e> Acesso em jun de 2018.

Widdows, D. e Ferraro, K. (2008). Semantic Vectors: a Scalable Open Source Package and Online Technology Management Application. LREC'08 Proceedings of the Sixth International Conference on Language Resources and Evaluation, n. January 2008.

Salles, S. de Britto (2010). "Questões discursivas: Cinco cuidados para uma boa resposta", https://educacao.uol.com.br/disciplinas/portugues/questoes-discursivas-cinco-cuidados-para-uma-boa-resposta.htm, Junho/2017.

Santana Junior, J. J. B., Pereira, D. M. V. G. e Lopes, J. E. (2008). Análise das habilidades cognitivas requeridas na Administração Pública Federal fundamentados na visão da Taxonomia de Bloom. Revista Contabilidade & Finanças, v. 19, n. 46, p.108-121.

Santos; Carlos Alves dos. Avaliação Automática de Questões Discursivas Usando LSA/ dos Santos. Tese (Doutorado) – Universidade Federal do Pará – UFPa Instituto de Tecnologia Programa de Pós-Graduação em Engenharia Elétrica. 2016

UFBA (2011). "Questões Discursivas em História: Interpretação e Comandos", http://www.portalmodulo.com.br/userfiles/Respodendo%20Discursivas(1).pdf, 2017.

Ziai, R., Ott, N. & Meurers, D. (2012). Short Answer Assessment: Establishing Links Between Research Strands. Proceedings of the Seventh Workshop on Building Educational Applications Using NLP. Association for Computational Linguistics.