Diversity Matters!
Proceedings of the European Distance and E-Learning Network 2017 Annual Conference
Jönköping, 13-16 June, 2017
ISBN 978-615-5511-18-9
ISSN: 2707-2819
doi: https://doi.org/10.38069/edenconf-2017-ac-0044

WRITING TO LEARN WITH AUTOMATED FEEDBACK THROUGH (LSA) LATENT SEMANTIC ANALYSIS: EXPERIENCES DEALING WITH DIVERSITY IN LARGE ONLINE COURSES

Miguel Santamaría Lancho, Mauro Hernández, Jose Maria Luzón Encabo, Guillermo Jorge-Botana, UNED, Spain

Abstract

The increasing demand for higher education and life-long training has induced a raising supply of online courses provided both by distance education institutions and conventional face to face universities. Simultaneously, public universities' budgets have been experiencing serious cuts, at least in Europe. Due to this shortage of human and material resources, large online courses usually face great challenges to provide an extremely diverse student community with quality formative assessment, specially the kind that offers rich and personalized feedback. Peer to peer assessment could partially address the problem, but involves its own shortcomings.

The act of writing has been identified as a high-impact learning tool across disciplines, and competence in writing has been shown to aid in access to higher education and retention. Writing to learn (WTL) is also a way to foster critical thinking and a suitable method to train soft skills such as analysis and synthesis abilities. These skills are the base for other complex learning methodologies such as PBL, case method, etc. WTL approach requires a regular feedback given by dedicated lecturers.

Consistent assessing of free-text answers is more difficult than we usually assume, specially, when addressing large or massive courses. Using multiple choice *objective* assessment appears an obvious alternative. However, the authors feel that this alternative shows serious shortcomings when aiming to produce outcomes based on written expression and complex analysis.

To face this dilemma, the authors decided to test an LSA-based automatic assessment tool developed by researchers of Developmental and Educational Psychology Department at UNED (Spanish National Distance Education University) named GRubric. The experience was launched in 2014-2015. By using GRubric, we provided automated formative and iterative feedback to our students for their open-ended questions (70-200 words). This allowed our students to improve their answers and practice writing skills, thus contributing both to better

organize concepts and to build knowledge. In this paper, we present the encouraging results of our first two experiences with UNED Business Degree students in 2014/15 and 2015/16.

Writing to learn (WTL)

The act of writing has been identified as a high-impact learning tool across disciplines, and efficacy in writing has been shown to aid in access and retention in higher education. Writing has also been shown to be effective in the promotion of learning and student success in relatively large enrolment face-to-face courses. Research suggests that writing instruction in online settings can provide enhanced learning experiences and opportunities for pedagogical reflection (Comer, Clark, & Canelas, 2014). The use of WTL can improve student understanding of contents and concepts; in addition, it can be an effective tool in student learning and engagement. Finally, WTL helps students to retain what they have to learn.

Furthermore, this approach promotes a deep learning also it is a suitable method to train soft skills such as critical thinking and the ability to analyze and synthesize (Forsman, 1985). These skills are at the base of other complex learning methodologies such PBL, case method, etc.

In spite of, evidence that writing can be an effective tool to promote student learning and engagement, writing-to-learn (WTL) practices are still not widely implemented, particularly at large online courses. One possible explanation is that WTL requires a regular feedback given by dedicated lecturers. Without such feedback, much of the learning potential of WTL is missed.

Giving feedback is one of the requirements to ensure the effectiveness of WTL. This feedback should be provided by teachers, lectures or experts in the subject. However, the increasing number of students and the subsequent workload make very difficult for university teachers to stick to this kind of exercise. Moreover, feedback makes possible the personalization of learning, fostering performance improvement and increasing motivation, as well. But what kind of feedback is demanded nowadays? Our students, as users of technologies, demand a quick and iterative feedback; for instance, they are accustomed to the trial and error method to learn how to handle technological devices and applications. Therefore, the challenge is how to give them quick, iterative and sustainable feedback when quality feedback is required, such as in WTL, and human instructors are not available, or not available enough.

An automated-assessment system for free-text short-answer questions (G-Rubric)

Automated Essays Assessment (AEA) has a long history. The development of technologies such as word processing and the Internet, encouraged the improvement of AEA systems. In addition, the advances experienced since the 1990's in computational technologies of natural language processing facilitated the analysis of morphology (word structure), syntax (sentence structure) and semantics (meaning). The analysis of content was carried out through lists of

ISBN 978-615-5511-18-9

keywords, synonyms and the analysis of the frequency with which certain terms appeared (Shermis & Burstein, 2003).

Recently, several new approaches have been explored, being Latent Semantic Analysis (LSA) one of the most promising developments. Research has burst in the last decade, with special focus on its application to education, although basically on small scale environments. Paradoxically, there is not so much research in Distance Education institutions, in spite of that massive numbers of students should have encouraged this field (Jorge-Botana et al., 2015). Concerns about plagiarism and identity-control issues have presumably hindered progress in this specific context, along with logistical issues related to access to computers at the examination place. At present, MOOCs represent, indeed, an obvious field for the implementation of this kind of application.

In general, according with previous research, AEA scoring tends to be accurate. Human and computer-assigned scores correlate around 0.80 to 0.85, with 40-60% perfect agreement and 90-100% adjacent agreement (human and computers scores within 1 point). Some AEA systems have become embedded within automated writing evaluation systems than assign scores along with feedback on errors and may include instructional scaffolds and learning management tools (Roscoe & McNamara, 2013).

LSA – What is it?

Latent semantic analysis (LSA) is based on the concept of vector space models. This means using linear algebra for allocating lexical units in an n-dimensional vector space. LSA is a set of different procedures by which a textual corpus, usually lemmatized and curated, is transformed into a semantic space. In a first step, this corpus is expressed into an occurrence matrix, which usually includes its terms as rows and paragraphs as columns. A second step is applied to this matrix which smoothes the asymmetries in word frequencies. The third step has made LSA famous which is applying to this matrix a dimension reduction technique by means of singular value decomposition (SVD) which provides a suitable space in which words and texts are represented in a few but relevant latent (with no meaning) dimensions. This space is very useful to represent expert and student answers and calculate similarities between them. The more similarity among student-expert answers, the higher score. But recently, some authors have developed a very promising procedure called inbuilt-rubric (Olmos, Jorge-Botana, León, & Escudero, 2014) which transforms the k first latent dimensions of the original space into non-latent dimensions. The k first dimensions no longer reflect latent knowledge, but reflect conceptual axes spread from relevant words of the academic topics. This is very useful to offer a conceptual feedback. The scores of the student answers in such k first dimensions indicates if the relevant concepts of the rubric are present in his answer. This technique has reached satisfactory results in real contexts (Olmos, Jorge-Botana, Luzón, Martín-Cordero, & León, 2016). This is just the procedure GRubric, the AEA of this study, uses.

ISBN 978-615-5511-18-9

For the Economic History teachers involved in this study, the most important characteristics of GRubric refer to its ability to provide the student with, at least, three different kinds of feedback for his/her answer to the short question posed: a numeric grade for content, an additional numerical grade for writing quality and a third, detailed graphic feedback which plots the score in each conceptual axe of the rubric. These scores, actually, are the scores in the first k-dimensions of the vector that represent the student answer. This is due to that inbuilt-rubric method imposes the meaning of every conceptual axis of the rubric to that k-dimensions of the space.

To do this, teachers had to provide/create two different types of inputs (which are the inputs to make GallitoAPI work):

1. *General texts for the corpus*: this is the raw material of the course (handbooks, reference texts, etc.), to be inserted on the corpus.
2. To generate the space from the corpus, all processes mentioned above are carried out through a specific program called Gallito Studio (Jorge-Botana, Olmos, & Barroso, 2013). Then, the resultant space, including inbuilt-rubric space, are upload to a specific API (Application Programming Interface) called GallitoAPI (www.gallitoapi.net) developed by researchers at UNED's Department of Developmental and Educational Psychology. The web interface for assessment of free-text was baptized as G-Rubric and we will usually refer to the whole system with this name, although it is important to retain than managing of the multi-vector semantic space, which is the heart of the system, is conducted via GallitoAPI. For our experience, we built a corpus on Economic History using six different World Economic History textbooks, all of them written in Spanish, and published in the last twenty years.

To accompany each question, we prepared a canon answer (or *golden text*) with which students' answers would be compared. A series of *conceptual axes* (three-five per question) were prepared for each question, made of a series of keywords that depict different regions of the semantic field the answer should cover. This golden text and *axes* were tested with actual students' answers taken from past exams in order to test the accuracy of the numerical grade and the graphic feedback drawn from conceptual axes. Several iterations were needed to reach acceptable objects for a trial with students. This material allows the system to process and assess free-text answers and provide students both with numerical grades for content and composition and a graphical feedback regarding conceptual axes. A web interface, named as G-Rubric, allows users to easily select questions and submit answers, and receive feedback almost immediately.

In order to help to understand how GRubrics works, we offer a sample of those activities proposed to our students.

Once the student registers in GRubric website and chooses the activity, he/she can write down/paste an answer. We have chosen an activity on the concept of Mercantilism.

ISBN 978-615-5511-18-9

Fist attempt, student's response:

> *"Mercantilism is a set of ideas and policies deployed in early modern Europe (16th, 17th and 18th centuries) aimed at strengthening the State through economic power, and specially focused on trade-balance surpluses and accumulation of precious metals (bullionism)."*

After submitting an answer, he/she receives the feedback that can be seen on the left side of the figure below. After examining this feedback, the student can review the earlier answer and make a new attempt adding, for instance, some new ideas about mercantilist policies (bold text in the second attempt).

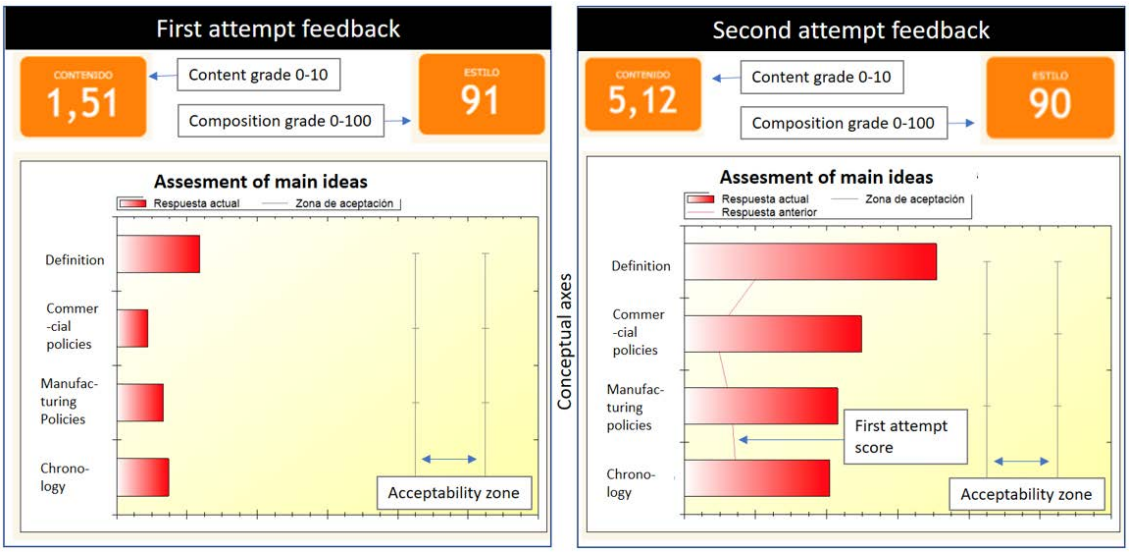


Figure 1. Screenshots of GRubric's feedback screen

Second attempt:

> *"Mercantilism is a set of ideas and policies deployed in early modern Europe (16th, 17th and 18th centuries) aimed at strengthening the State through economic power, and specially focused on trade-balance surpluses and accumulation of precious metals (bullionism). Amongst mercantilist polices, some outstand, i.e. those focused on attaining surpluses in trade balance through tariff protection, prohibition of exports of gold, silver and raw materials, creation of chartered trade companies, navigation acts and commercial monopolies."*

A new feedback is produced, as seen on the right side of the Figure 1. Then, the student can try again using the new feedback to improve his/her answer.

Experiences carried out in 2015-2016: description and main results

It is important to point out that the trials carried out along the last two years were focused on providing formative assessment. Our goal was to promote deep learning through iterative

ISBN 978-615-5511-18-9

feedback, not just grading student´s assignments. GRubric offers two main advantages regarding formative assessment: it allows as many attempts as lecturers set and gives the students immediate rich feedback. All trials have been conducted with first year Business Administration Degree students.

First experience with GRubric (May 2015)

Whit this first experience we had two goals: first to determine the efficacy of GRubric to promote learning and second to establish its reliability to mark student's assignments. To develop this first trial, we asked for volunteers between our students and offer them a little reward (adding 0.25 point to their final mark). We got 132 volunteers and we split them randomly into 3 groups establishing different conditions for each group. Group 1 received rich feedback, both numerical and graphical, and had 6 attempts to answer. Group 2 received poor feedback (only numerical) and had also six attempts. Finally, Group 3 was the control group and received poor feedback and only one attempt per object was allowed.

The students taking part in the trial would answer five short open questions (between 70 and 200 words), very similar to those they would find in their final exam. For each question, the student got a set of instructions referring to the number of words expected to write, how to use the tool to answer, and guidance for using the received feedback. Groups 1 and 2 could use six attempts to improve their answers according to the received feedback. Each student could decide how many attempts he would make. The difference between the worst and the best mark achieved in each of the activities was used to measure the learning improvement of each student. In addition, a questionnaire was used to measure student´s agreement with the grades assigned by GRubric to their answers.

As can be seen in Table 1, in general, there was a learning improvement for Group 1 as well as for Group 2. Also, the difference between highest and lowest grades was higher for the Group 1, which received rich feedback.

Table 1: Trial 2015. Improved learning indicators

		Average Grade GRubric (/10)			Difference between max-min grade		
	Item	G1	G2	G3	G1	G2	G3[(1)]
1	Demographics regimes	6.9	6.5	6.4	0.52	0.69	0
2	Consequences of the Neolitic Revolution	6.5	5.9	5.6	1.06	0.95	0
3	European agrarian economies during Middle Ages	6.2	7.4	5.5	1.10	0.78	0
4	Mercantilism	7.7	7.5	6.6	1.95	1.15	0
5	(Final) Colonial Commerce[(2)]	6.2	6.3	6.1	0	0	0

[(1)] G3 was the control group and had only one attempt per item, then there was no option to improve
[(2)] For Item 5 only one attempt was allowed.

As for student's agreement with the grades received, as we can see at Figure 1, it marked quite well.

ISBN 978-615-5511-18-9

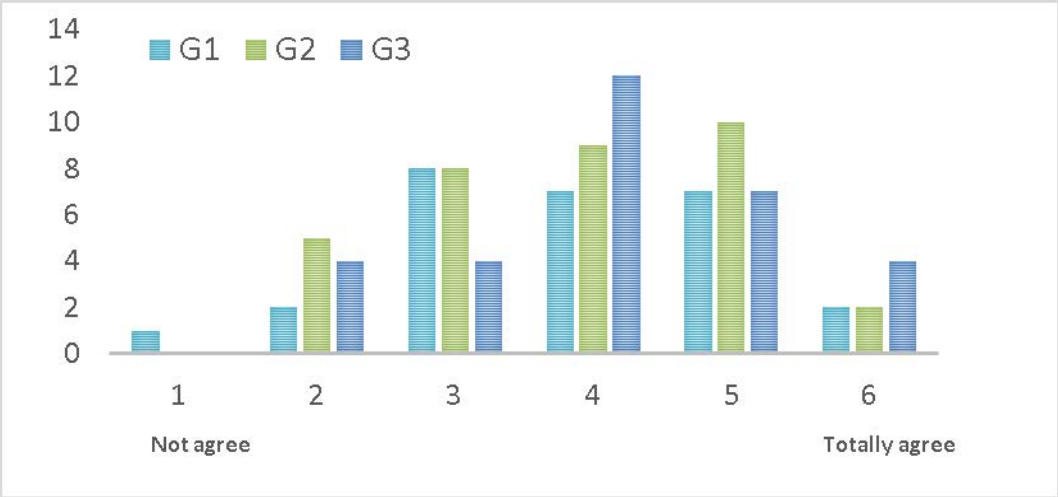


Figure 2. Student´s agreement with the grades received

Second experience with GRubric (April-May 2016)

The goal for this second trial was to improve the design of GRubric objects to fostering learning and increasing student satisfaction. In order to carry out this second trial, we devoted time to set up new questions, increasing from five to seven the number of objects offered to the students. To increase the number of volunteers the reward was upgraded from 0.25 to 1 point. This reward was associated with the number of attempts performed, rather than with the grades produced by GRubric, because after the first experience we discovered that learning improved after several attempts at answering.

According to data in Table 2, the average grades obtained were satisfactory. It should be taken into account that we had recommended to the students that they should review the textbook before producing an answer. As we can see, after students accessed to feedback they were able, on average, to improve their marks in the following attempts.

It is also worth to note that the best students were able to obtain high scores, very close to those of the *golden essay* produced by the lecturer and used by the system as a reference to mark students' submissions.

Table 2: Trial 2016. Student's scores by item

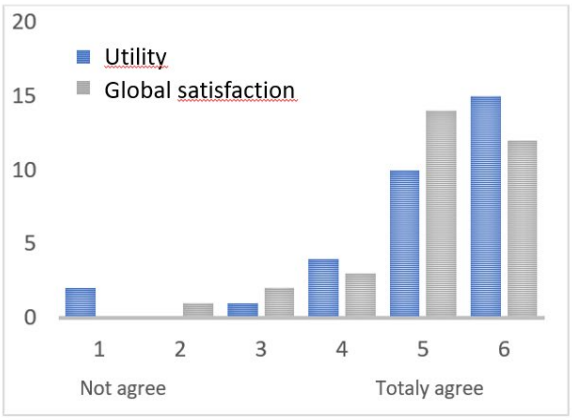
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Average
Lowest grade	6.19	4.51	5.10	4.95	5.39	5.02	5.66	5.19
Highest grade	7.41	5.53	6.12	5.81	6.74	6.29	6.78	6.31
Difference in points	1.22	1.02	1.02	0.86	1.34	1.27	1.12	1.12
Difference %	19.67	22.62	19.95	17.45	24.93	25.36	19.80	21.66

To analyze learning improvement (i.e. *learning*) we used the difference between the lowest and highest grade obtained by students. Table 2 shows the difference by item, both in absolute term and as a percentage. A 21.6% improvement average could be considered as remarkable, given that only three attempts were allowed. The different degree of improvement by item

ISBN 978-615-5511-18-9

could be a consequence of different factors such as the quality of the item design, difficulty of the item, etc.

To conclude the analysis of this second trial, we would like to point out some results of the satisfaction questionnaire that students completed after the experience (Figures 3 and 4).



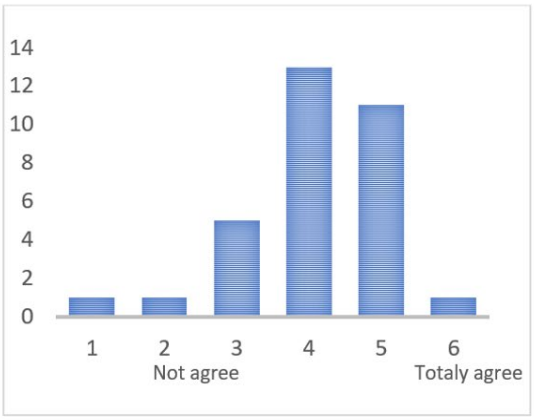


Figure 3. Trial 2016: Utility ad satisfaction with GRubric app
Figure 4. Student's agreement with grade obtained

According to Figure 2, students considered the experience very useful and believed that they were better prepared for the final exam. Global satisfaction was also high.

Regarding the student's agreement with the grades received, as we can observe in Figure 3, it could be said that it was quite satisfactory.

Conclusions

Some conclusions can be drawn from our experience:

1. Automated-assessment software such as Gallito–GRubric is currently mature enough to be used with students obtaining quite satisfactory results in terms of acceptable accuracy. Results in terms of students' satisfaction are also encouraging. Developments in this area, especially with LSA-based systems, will probably be added to our teaching toolbox in the near future.
2. This kind of systems is particularly apt and useful for on-line teaching, especially in massive courses such as MOOC, in which the great number of students often poses serious challenges to the scarce teacher's hours. Nevertheless, they show also great potential for face-to-face or mixed teaching at any level.
3. The experience of adapting such a system to assess open-ended questions to Economic History proved reasonably affordable in terms of time and effort invested. Learning to work with GRubric was also easy for students, although there are some indications that mastering the system – and especially fully understand graphic feedback – could take them a little more than expected.

ISBN 978-615-5511-18-9

4. The trial's results seem to point out that interacting with GRubric can improve learning by giving detailed feedback: (a) encourages devoting more time to the task; (b) increases *earnings* in the quality of answers; (c) increases motivation to work on activities (d) helps students to achieve better final answers. In this sense, it may soon become a viable tool for formative assessment.
5. Although it requires further research, the accuracy of GRubric, both as perceived by teachers and students, offers a great potential for its use in summative assessment, as well.

In the near future, automated assessment systems will be part of the teacher's toolbox, as Virtual Learning Environments are today. LSA-based systems such as GRubric are a solid candidate to a leading role in that process.

References

1. Comer, D. K., Clark, C. R., & Canelas, D. A. (2014). Writing to learn and learning to write across the disciplines: Peer-to-peer writing in introductory-level MOOCs. *The International Review of Research in Open and Distributed Learning, 15*(5). Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/1850
2. Forsman, S. (1985). Writing to learn means learning to think. In A. Ruggles Gere (Ed.), *Roots in the Sawdust* (pp. 162–174).
3. Haley, D. T., Thomas, P., De Roeck, A., & Petre, M. (2005). *A research taxonomy for latent semantic analysis-based educational applications*. Technical Report no. 2005/ 09. Open University.
4. Haley, D. T., Thomas, P., Petre, P., & De Roeck, A. (2007). *Seeing the whole picture: Comparing computer assisted assessment systems using LSA-based systems as an example*. Technical Report Number 2007/07. Open University.
5. Jorge-Botana, G., Olmos, R., & Barroso, A. (2013, July). Gallito 2.0: A natural language processing tool to support research on discourse. *Proceedings of the 13th Annual Meeting of the Society for Text and Discourse, Valencia, Spain*.
6. Jorge-Botana, G., Leon, J. A., Olmos, R., & Escudero, I. (2010). Latent semantic analysis parameters for essay evaluation using small-scale corpora. *Journal of Quantitative Linguistics, 17*(1), 1-29.
7. Jorge-Botana, G., Luzón, J. M, Gómez-Veiga, I., & Martín-Cordero, J.(2015): Automated LSA assessment of summaries in Distance Education: Some variables to be considered. *Journal of Educational Computing Research, 52*, 341-364.
8. Olmos, R., Jorge-Botana, G., León, J. A., & Escudero, I. (2014).Transforming Selected Concepts into Dimensions in Latent Semantic Analysis. *Discourse Processes, 51*(5-6), 494-510.

ISBN 978-615-5511-18-9

9. Olmos, R., Jorge-Botana, G., Luzón, J. M., Martín-Cordero, J. I., & León, J. A. (2016). Transforming LSA space dimensions into a rubric for an automatic assessment and feedback system. *Information Processing & Management, 52*(3), 359-373.
10. Roscoe, R. D., & McNamara, D. S. (2013). Writing pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology, 105*(4), 1010.
11. Shermis, M. D., & Burstein, J. (Eds.) (2003). *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
12. Tonta, Y., & Darvish, H. R. (2010). Diffusion of latent semantic analysis as a research tool: A social network analysis approach. *Journal of Informetrics, 4*(2), 166-174.

ISBN 978-615-5511-18-9