Escribir para aprender: evaluación automática de respuestas abiertas con G-Rubric

MIGUEL SANTAMARÍA LANCHO, JOSÉ MARÍA LUZÓN ENCABO, MAURO HERNÁNDEZ BENÍTEZ
y GUILLERMO DE JORGE BOTANA
Universidad Nacional de Educación a Distancia (UNED)
msantamaria@cee.uned.es

Resumen. **El incremento de la demanda de formación en línea junto con los recortes experimentados en los últimos años han contribuido a empobrecer el feedback que reciben los estudiantes y a concentrar la evaluación en pruebas objetivas.**
"Escribir para aprender" es un método que impulsa el desarrollo del pensamiento crítico, la capacidad de síntesis y de análisis. Lo cual está en la base de otras metodologías más complejas como el ABP, pero utilizar el escribir para aprender como herramienta de aprendizaje requiere dar feedback manual.
Para hacer posible la utilización del "escribir para aprender" y poder facilitar el feedback requerido en una asignatura con muchos estudiantes, el equipo docente de Historia Económica ha comenzado a utilizar una herramienta tecnológica desarrollada en la UNED, por el departamento de Psicología Evolutiva y de la Educación. Dicha herramienta está basada en la utilización de técnicas de Análisis Semántico Latente. Esta herramienta es capaz de facilitar feedback cuando responden a preguntas de respuesta abierta. Esto permite al estudiante mejorar su respuesta de manera iterativa.

Palabras clave: *escribir para aprender, feedback enriquecido, corrección automática de respuestas abiertas, desarrollo competencias tranvsersales.*

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 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 G-Rubric. The experience was launched in 2015-2016. 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 experience with UNED Business Degree students in 2015/16.

Keywords: *writing to learn, rich feedback, automathic feedback open-ended questions, transferable skills development.*

INTRODUCCIÓN

El elevado número de estudiantes matriculados en las asignaturas de los primeros cursos de la UNED ha tenido como consecuencia, en el ámbito de la evaluación de los aprendizajes, el uso y abuso de pruebas objetivas (test). La mayor parte de los docentes son conscientes de las limitaciones que ello supone para hacer una evaluación más allá del mero reconocimiento de informaciones.

En asignaturas del ámbito de las Ciencias Sociales esto supone un hándicap añadido. El "feedback" queda reducido a la autoevaluación, pues el equipo docente puede facilitar a los estudiantes rúbricas u orientaciones para guiar la autoevaluación. Lo deseable sería poder ofrecer una evaluación externa facilitada por un tutor.

Por otro lado, los integrantes de esta red consideran que la escritura no es solo relevante para desarrollar las habilidades relacionadas con la comunicación escrita, sino que guarda relación directa con el proceso de construcción de aprendizajes por parte de los estudiantes. Cuando el estudiante es capaz de explicar con sus propias palabras lo aprendido, tenemos una señal de que efectivamente se ha producido aprendizaje.

Con esta experiencia pretendemos analizar las posibilidades que ofrece un conjunto de herramientas informáticas desarrolladas en la UNED que facilitan, mediante tecnologías basadas en el procesamiento de lenguaje natural, la corrección automática de pruebas de respuesta abierta. En esta primera etapa nos centraremos en su utilización como herramienta de apoyo a la evaluación formativa.

La experiencia se ha llevado a cabo en la asignatura Historia Económica del Grado de ADE en la que durante el curso 2015-16 han estado matriculados 3.800 estudiantes. Pese al número de estudiantes la asignatura tiene como examen final, una prueba integrada por cinco preguntas cortas, similares a las utilizadas en esta experiencia, una parte práctica y una prueba tipo test con 10 preguntas.

Con esta experiencia también se pretende encarar el problema que supone la discrepancia entre correctores en la corrección manual. Cuando se han analizado las diferencias entre las calificaciones otorgadas a un mismo examen por diferentes profesores de la asignatura se ha comprobado que estas diferencias existen y no son menores.

Objetivos del proyecto

- Probar la utilidad de la aplicación GRubric para corregir y dar feedback en pruebas de respuesta abierta en actividades de evaluación formativa.

- Analizar el impacto que tiene este feedback en la mejora del desempeño. A partir de la información facilitada por el feedback el estudiante puede elaborar nuevas respuestas, que serán de nuevo evaluadas. Se trata de ver si con los sucesivos intentos se produce una mejora en el rendimiento a través de la mejora de la calificación obtenida.

METODOLOGÍA

El equipo docente preparó y calibró 7 ítems o preguntas a los que deberían responder los estudiantes. Estos ítems eran similares a las preguntas cortas que se incluyen en el examen final de la asignatura.

Anuncio de la experiencia: solicitud de voluntarios.

Para llevar a cabo la experiencia se creó un foro en el curso virtual de la asignatura a través del cual se invitó a los estudiantes a participar en la experiencia. A través de este mensaje se les invitaba a inscribirse para realizar la experiencia.

En la experiencia han participado 105 estudiantes:

Incentivo por participar en la experiencia

Como incentivo a la participación se ofreció a los estudiantes la posibilidad de sustituir la realización de la segunda PEC de la asignatura por la participación en esta experiencia. Esto equivalía a incrementar en un punto la nota final, si la nota de la PEC fuese de 10 puntos.

Para los estudiantes que optasen por conseguir puntuación para su nota final participando en la experiencia de GRubric, su puntuación no estaría relacionada con la nota que obtuviesen tras contestar las preguntas en GRubric, sino con el número de intentos de respuesta que realizasen. Lo que queríamos incentivar era el uso de la herramienta repitiendo los intentos de respuesta para valorar la mejora en el rendimiento que se conseguía tras intentos sucesivos. El número máximo de intentos que podía hacer un estudiante eran 21, pues se le presentaban 7 preguntas con tres intentos por cada una de ella. La escala de puntuación en función de los intentos realizados fue la siguiente:

* 15 intentos: 5 puntos sobre 10

* 17 intentos: 7 puntos sobre 10

* 19 intentos: 9 puntos sobre 10

* 20 intentos: 10 puntos sobre 10

RESULTADOS PRINCIPALES

En la tabla 1 se muestran los datos básicos respecto a la cómo los estudiantes respondieron a cada una de las preguntas en términos de número de intentos, notas promedio obtenidas por el conjunto de estudiantes, así como notas máximas conseguidas de manera individual.

Tabla 1. *Resultados de la experiencia para cada pregunta (número de intentos, nota media y nota máxima)*

Pregunta	Número de intentos	Promedio de puntuación	Puntuación máxima
Pregunta 1	248	6,87	9,67
Pregunta 2	216	5,03	9,19
Pregunta 3	177	5,68	9,57
Pregunta 4	162	5,38	9,06
Pregunta 5	158	6,22	9,29
Pregunta 6	135	5,73	9,38
Pregunta 7	128	6,40	9,33
Total general	**1224**	**5,9**	**9,67**

Como puede verse, los resultados en términos de notas en cada actividad son razonables, y parten de medias ya elevadas, aunque es importante tener en cuenta que una vez conocido el enunciado de la pregunta de cada actividad se les indicaba que consultaran el manual para tratar de dar una primera respuesta satisfactoria, que debía mejorar a partir del feedback recibido en el primer y segundo intentos (en el tercero y último recibían feedback, pero ya no había posibilidad de rectificar). Quizá el dato más importante lo dan las notas máximas absolutas, que en todas las actividades el 9 (hay que aclarar, por otro lado, que la puntuación obtenida con la llamada "repuesta de oro" no es nunca un 10, debido a una función de ajuste del sistema.

Los resultados permiten afirmar que la aplicación califica razonablemente bien (las respuestas mínimas están muy cerca del aprobado) y es sensible a las mejoras (aunque no en todas las actividades por igual).

Adicionalmente, la distribución de notas por preguntas (ilustración 1) nos permite testar el grado de dificultad/calidad de cada una, lo que puede resultar muy útil para la programación de las actividades

El análisis del aprendizaje, tomando como indicador diferencia entre el mejor y el peor intento en cada actividad (tabla 2), no permite apreciar una mejora sustancial en términos absolutos (promedio: 1,4 puntos sobre 10), ni tampoco una mejora creciente de las primeras actividades a las últimas, aunque sí diferencias –aunque no muy grandes: de un mínimo de 1,2 a un máximo de 1,7-- en la mejora en las distintas actividades, seguramente explicables por diferencias en el diseño de los objetos: los mejor diseñados en la pregunta, respuesta de oro y ejes conceptuales serían en principio los que daban mejoras más altas. Pero si lo medimos en término relativos (porcentaje de mejora de la mejor nota respecto a la peor), las cifras dejan de parecer insignificantes: en solo tres intentos se produce una mejora de puntuación del 40% (promedio), que en algunas actividades llega casi al 70% (tabla 3)

Tabla 2. *Ensayo 2016. Aprendizaje. Diferencia entre el mejor y el peor intento, en término absolutos (puntos de nota) y relativos (% de mejora sobre peor respuesta). Promedios*

Actividad	1	2	3	4	5	6	7	Todos
Mejora promedio en puntos de nota	1,5	1,3	1,3	1,2	1,7	1,6	1,3	1,4
Mejora promedio en %	42	34,5	39	31	67	40,7	28,7	40,4

Análisis de la relación del desempeño de los estudiantes en la experiencia y en la prueba presencial.

En primer lugar, hemos de admitir que existe un sesgo de autoselección, pues fueron los mejores estudiantes los que más se involucraron en la actividad.

En la tabla 3 podemos ver el comportamiento del grupo que participó en GRubric frente al conjunto de estudiantes de la asignatura.

Tabla 3. Rendimiento comparado de los estuiantes que participaron en GRubric y el resto de matriculados

Indicador		
	Grupo Grubric	Total estudiantes
Presentación a examen	41 (87,2%)	891 (32,1%)
Presentados a PECs*	47 (100%)	467 (52%)
Nota media PECs	9,7	6,6
Nota media examen	5,2	3,46 (-21,8%)
Nota media test/2	1,2	0,82 (-18,4%)
Nota cortas/5	2,5	1,58 (-19,2%)
Nota comentario/3	1,4	1,06 (-12,8%)

Finalmente, los estudiantes que participaron en la experiencia de GRubric se presentaron más al examen final en la convocatoria de junio y tuvieron unas mejores calificaciones tanto en la calificación global como en cada una de las partes del examen (test, preguntas cortas, actividad práctica).

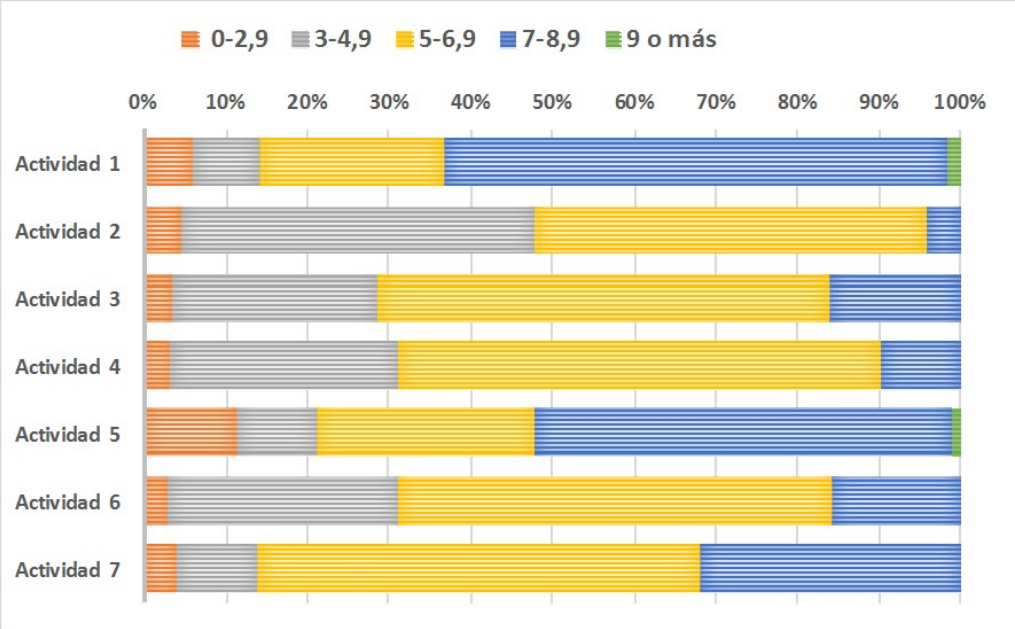

Figura 1. Distribución de las calificaciones por cada pregunta

CONCLUSIONES

a) La evaluación "humana" de ejercicios de texto libre presenta problemas a los que los profesores solemos cerrar los ojos –porque sería atrevido decir que no somos conscientes de ellos-- solemos ignorar y que merece la pena investigar sistemáticamente. Con los datos limitados aquí expuestos se puede afirmar que incluso ante un mismo examen (véanse los casos de doble corrección) hay diferencias sustanciales en la nota, y eso pese a la presencia de elementos objetivos de homogeneización. Pero incluso la corrección ordinaria muestra sesgos visibles y sistemáticos achacables a las preferencias de cada profesor, sin que haya que dar por hecho que un mismo profesor califica siempre de forma estable. Dicho más crudamente: la evaluación tradicional de textos libres no es suficientemente fiable y coherente, especialmente dadas las crecientes demandas de los estudiantes en este sentido.

b) Como alternativa a una evaluación empobrecida basada en tests de elección múltiple, el software de evaluación automática como Gallito-GRubric (y seguramente otros que existan o se estén desarrollando) está lo suficientemente maduro para pasar a la fase de pruebas con estudiantes reales. Desde luego, es así en lo que se refiere a la evaluación formativa.

c) Estas herramientas son particularmente útiles en la enseñanza on-line o semipresencial, ya sea en cursos ordinarios o MOOCs, donde el número masivo de estudiantes impiden recurrir a los costosos servicios de un profesorado escaso y cargado de obligaciones. Pero también presentan un potencial importante en la enseñanza presencial o mixta en cualquier nivel.

d) Entre las virtudes de estas aplicaciones está el que proporcionan una experiencia "gamificada" (similar a un juego),con feedbacks inmediatos y ya relativamente ricos, que proporcionan refuerzos positivos inmediatos a los usuarios, mejorando así su adherencia a la actividad, y por tanto el aprendizaje.

e) Nuestra experiencia en adaptar un sistema de este tipo a preguntas cortas de texto libre de historia económica ha demostrado ser razonablemente asequible en tiempo y esfuerzos. El aprendizaje de G-Rubric por parte de los estudiantes también parece ser poco costoso, aunque hay indicios de que llegar a dominar el juego puede llegar a costarles más de lo que esperarían en principio.

f) Aunque seguimos con este proceso de pruebas de Gallito-G-Rubric, tanto en esta asignatura como en otras de la UNED, creemos que la percepción tanto de los profesores como de los estudiantes revelan un potencial importante de cara a la evaluación sumativa.

g) Vistos los problemas reseñados de la evaluación humana, cabe considerar el uso de sistemas basados en LSA como mecanismo de control o refuerzo de ésta. Así, análogamente a lo que se hace con el software de traducción automática, podríamos dejar que estas aplicaciones nos proporcionaran una nota en borrador, que el docente podría luego refinar con una lectura "humana".

En suma, creemos que los sistemas de evaluación automática se incorporarán en no más de una década a la caja de herramientas del profesorado, también en la universidad. En este proceso, los sistemas basados en LSA, como Gallito/G-Rubric, son un candidato sólido a desempeñar un papel protagonista en el proceso. Con suerte, nos ahorrarán mucho trabajo mecánico de corrección, liberando tiempo para una docencia de más valor añadido. En el peor de los casos, nos permitirán que los estudiantes practiquen una evaluación formativa intensiva y de cierta calidad. Y seguramente nos ayudarán a mejorar nuestros procedimientos en la evaluación, mitigando la inestabilidad de sus resultados.

BIBLIOGRAFÍA

Bernardos Sanz, J.U., Hernández, M. & Santamaría Lancho, M. (2014) Historia Económica. UNED. Madrid.

Biggs.J. & Tang, C. (2007) Teaching for Quality Learning at University. MacGraw Hill-Open University. Londres.

Comín, F., Hernández, M. & Llopis, E. (Eds.) (2005) Historia Económica Mundial (ss. X-XX). Crítica. Barcelona.

Comín, F. (2011) Historia económica mundial: de los orígenes a la actualidad. Alianza. Madrid.

Dunn, L., Morgan, C., O'Reilly, M. & Parry, S. (2003) The Student Assessment Handbook: New Directions in Traditional and Online Assessment. Routledge Falmer. Londres-Nueva York.

Fahmy Yousef, A.M., Wahid, U., Amine Chatti, M., Schroeder, U. y Wosnitza, M. (2015) " The Effect of Peer Assessment Rubrics on Learners' Satisfaction and Performance Within a Blended MOOC Environment ", CSEDU (2), pp. 148-159 [Consulta 20 mayo 2016]. Disponible en: https://www.researchgate.net/profile/Ahmed_Mohamed_Fahmy_Yousef/publication/278675891_The_Effect_of_Peer_Assessment_Rubrics_on_Learners'_Satisfaction_and_Performance_Within_a_Blended_MOOC_Environment/links/5582d10408ae6cf036c2f83b.pdf.

Hernández Benítez, M. y Bernardos Sanz, J.U. (2014) "Cursos virtuales ¿qué hay ahí dentro?", XI Encuentro de Didáctica de la Historia Económica, Santiago de Compostela, 26 y 27 de junio de 2014, [Consulta 20 mayo 2016]. Disponible en: http://www.usc.es/export/sites/default/es/congresos/xiedhe/papers/S4_8_Hernandez_Bernardos_TC.pdf

Hernández, M., Jorge-Botana, G., Luzón, J.M. y Santamaría Lancho, M. (2015) "Corrección automática de texto libre vs. corrección humana: ¿Qué o quién lo hace mejor?", Comunicación presentada a la XVII Reunión de Economía Mundial, Oviedo (3-5, junio 2015), Asociación de Economía Mundial- Universidad de Oviedo

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.

Kahneman, D. (2013) Pensar rápido, pensar despacio. DeBolsillo. Barcelona.

Mateos Royo, J.A. (2014) Historia económica mundial. Gráficas Huesca. Huesca.

Olmos, R., Jorge-Botana, G., León, J.A, Escudero, I. (2014)"Transforming Selected Concepts Into Dimensions in Latent Semantic Analysis". Discourse Processes Vol. 51, Num. 5-6: 494-510

Olmos, R., Jorge-Botana, G., Luzón, J.M., Martín, 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), pp. 359-373.doi:10.1016/j.ipm.2015.12.002 [ISSN: 0306-4573]

Palafox, J.A. (Ed.) (2014) Los tiempos cambian. Historia de la economía. Tirant Universitat. Valencia.

Scouller, K. (1998) "The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay", Higher Education, 35(4), pp 453-472

Shermis, M. D. & Burstein, J. (Eds.) (2003) Automated essay scoring: A cross-disciplinary perspective. Lawrence Erlbaum Associates, Inc. Hillsdale, (NJ)

Simón Segura, F. (1996) Manual de historia económica mundial y de España. CERA. Madrid.

Tascón Fernández, J. & Misael Arturo López Zapico (2012) Historia Económica Mundial. Una visión eurocéntrica de la actividad económica, del Neolítico al siglo XXI. Biblioteca Nueva. Madrid.

Valenti, S. Neri, F. & Cucchiarelli, R. (2003) "An overview of current research on automated essay grading", Journal of Information Technology Education, 2, 319-330.

Wakeford, R. (2003)" Principles of student assessment "in Fry, H. Ketteridge, S. & Marshall, S. (Eds.) (2003) A handbook for teaching & learning in higher education. Second edition, Kogan-Page. Sterling (VA): 42-61.

Werbach, K. y Hunter, D. (2012) For the Win: How Game Thinking Can Revolutionize Your Business, Wharton Digital Press, Philadelphia (PA).