G-Rubric: una aplicación para corrección automática de preguntas abiertas. Primer balance de su utilización

G-Rubric: an application for automatic assessment of free-text questions: first outcome analysis

Mauro Hernández
UNED

Miguel Santamaría Lancho
UNED

Resumen:

Evaluar de forma consistente las preguntas de respuesta libre resulta más difícil de lo damos por hecho. La evidencia resultantes de la experiencia de calificar/evaluar un elevado número de exámenes en la UNED hace aflorar alguna de estas dificultades. Frente a ellas, la alternativa más obvia es evaluar mediante preguntas "objetivas" de tipo test. Pero para los autores, esta alternativa resulta insuficiente para fomentar el desarrollo de aptitudes para la expresión y el análisis escrito complejo.
Por ello, nos animamos a tratar de abordar esta contradicción G-Rubric, una herramienta de evaluación automática basada en LSA desarrollada en el departamento de Psicología Evolutiva y de la Educación en la. En el curso 2014/2015 comenzamos las pruebas, que permiten facilitar a los estudiantes *feedback* formativo cuando éstos responden, a través de una aplicación web, a preguntas abiertas. Así, el estudiante puede mejorar sus respuestas y sobre todo practicar la expresión escrita, lo que ayuda tanto a organizar sus ideas como a apropiarse de los conocimientos.
La aplicación se basa en el procesamiento de lenguaje natural. Su utilización requiere la creación de un corpus semántico propio de la asignatura, para la cual hemos empleado seis manuales de Historia Económica obra de profesores españoles. Lo que se presenta aquí son los alentadores resultados de los primeros ensayos con esta herramienta, en los cursos 2014/15 y 2015/16, con estudiantes del grado de ADE en la UNED.

Palabras clave: LSA, evaluación automática, historia económica, sistemas expertos, GRubric

Abstract:

Consistent assessing of free-text answers is harder than we usually tend to assume. Evidence from assessing/grading big numbers of final tests at Spain's UNED (Open University) reveals some of these difficulties. Using multiple choice "objective" assessment appears an obvious alternative. But the authors this

alternative shows serious shortcomings at producing outcomes based on written expression and complex analysis.
To face this dilema, the authors decided to test an LSA-based automatic assessment tool developped by teachers of the Evolutionary and Educational Psichology ate UNED: G-Rubric. Tests were launched in 2014/2015, facilitating our students formative *feedback* for free-text answers to short questions displayed on a web application. Ths allows our students to improve ther answers and practice with written expression, thus contributing both to organizing concepts and to appropriate knowledge.
This application is based on natural language assessment. It requires the creation of a semantic corpus for the subject-matter, which we built from six Economic History textbooks written in Spanish. What we present here are the encouraging results of first tests with this app conducted in 2014/15 and 2015/16, with Business Degree (ADE) students at UNED (Universidad Nacional de Educación a Distancia).

Keywords: LSA, automatic assessment, economic history, expert systems, GRubric™.

Introducción

Estamos perfectamente acostumbrados a los sistemas de corrección automática, a través de escáneres ópticos que registran la respuesta en formularios que deben cumplimentarse rellenando distintas casillas: ese es el procedimiento estándar por ejemplo para exámenes masivos como los del MIR para acceder al especialidades médicas, o en la parte teórica del carné de conducir. La traducción de este tipo de cuestionarios a formatos web con corrección automática ha difundido aún más este tipo de pruebas, que se emplean para un sinfín de usos, desde encuestas de calidad, hasta meros tests de entretenimiento. La popularización de los cursos o aulas virtuales en la enseñanza (VLE por sus siglas en inglés) han contribuido aún más a esta difusión: prácticamente todas las plataformas de VLE, empezando por los *moodle* que parece el estándar en la universidad española (Hernández y Bernardos, 2014), incorporan este tipo de herramientas que permiten introducir cuestionarios de respuesta múltiple con diversos criterios de evaluación y diversas opciones en cuanto al *feedback* que se proporciona a los estudiantes.

Este tipo de herramientas de evaluación, basadas en preguntas de elección múltiple (comúnmente llamadas tipo test) presentan resultan particularmente útiles en entornos donde los estudiantes se cuentan por centenares, o incluso por millares, como pueden ser los *MOOCs* (Massive Online Open Courses), o en menor escala, en universidades a distancia como la UNED, en cuyas asignaturas de primer curso es frecuente contar con cientos de estudiantes –unos 3.000 en ADE o 1.000 en Economía, atendidos por cuatro profesores. Pero incluso con números menores de estudiantes, los exámenes tipo test resultan atractivos a la hora de calificar pruebas intermedias de evaluación continua o hasta exámenes finales. La rapidez y la "objetividad" en la corrección aparecen de cara a los alumnos como grandes ventajas de este tipo de pruebas. Con todo, presentan importantes limitaciones – en especial para la evaluación de aprendizaje profundo-- que reducen su aplicación en la mayoría de las universidades (Scouller, 1998).

De hecho, ha sido el boom de los *MOOCs* en los últimos años, por más que ahora parezca menos boyante, el que ha llevado a buscar estrategias de evaluación para grandes números de alumnos que sean a la vez más complejas y ofrezcan resultados más matizados que el simple test de elección múltiple. La necesidad de proporcionar un *feedback* rico para las tareas de los estudiantes, y hacerlo de forma rápida y a la vez fiable, ha llevado a explorar nuevos métodos de evaluación, algunos tan prometedores como la evaluación por pares basada en rúbricas (Yousef et al. 2015). Sin embargo, las posibilidades de sistemas automatizados de evaluación de texto libre, como el que presentamos aquí, no parecen haber gozado de tanta fortuna, pese a que sus ventajas se extienden incluso a entornos en los que el número de alumnos es menor.

Lo que presentamos aquí son los primeros resultados de las pruebas con uno de estos sistemas de evaluación automática de texto libre en una asignatura de Historia Económica en la UNED. La herramienta que empleamos, denominada G-Rubric™, es un sistema de evaluación automática de texto libre basado en

análisis de semántica latente (LSA), diseñado por un equipo de investigadores del Departamento de Psicología Evolutiva y de la Educación de la UNED. Tras someterla a prueba en el curso 2014/15 con un grupo de alumnos de primer curso del grado de ADE en la UNED, a partir de cinco actividades de respuesta corta (entre 75-200 palabras) relativas a la asignatura de Historia Económica, de nuevo hemos procedido a realizar una prueba en el curso 2015/16 con un grupo similar de alumnos. En ambos casos, nos hemos limitado a poner a prueba las posibilidades de la aplicación de cara a la evaluación formativa –no sumativa, es decir, las notas no se tienen en cuenta--, los resultados en ambos casos parecen respaldar la utilidad de esta herramienta, tanto en términos de fiabilidad de la evaluación como de satisfacción de los estudiantes, hasta el punto de llevarnos a considerar la posibilidad de aplicar G-Rubric a la evaluación sumativa ordinaria. Es decir, emplearlo para calificar a los estudiantes.

Por desgracia, los resultados en términos de aprendizaje como resultado del uso de la herramienta no son lo bastante sólidos con la prueba de 2014/15, ni el diseño de la prueba permite evitar en la del curso 2015/16 eliminar la influencia del sesgo de autoselección en el grupo de estudio. Por otro lado, las fechas en las que debe entregarse esta comunicación (mayo 2016) impiden comparar los resultados de la actividad con las notas finales en la asignatura. De este modo, las conclusiones serán más provisionales que nos gustaría, pero en todo caso, como decimos, prometedoras.

El texto comienza por examinar algunos problemas ligados a la evaluación humana frente a la automática, como introducción para defender la utilidad de esta última. Después se ofrece una introducción al funcionamiento de los sistema de evaluación basados en LSA, y en concreto de G-Rubric™, y finalmente examina los resultados de las pruebas llevadas a cabo con los estudiantes de la UNED en estos dos cursos. Finalmente, las conclusiones abordan la discusión de estos resultados, que podrían arrojar dudas sobre las ventajas de la evaluación humana frente a la automática.

1. Problemas de la evaluación por humanos

Posiblemente todos los profesores hemos oído alguna vez las quejas de alumnos sobre la "subjetividad" (o injusticia, si el tono sube) de la calificación de ejercicios de respuesta libre, tipo exámenes de desarrollo o comentarios de material práctico, e incluso preguntas de respuesta corta. Aunque solemos replicar a estas alegaciones con diversos argumentos, muchos de nosotros sabemos, en nuestro fuero interno, que algo de razón hay en estas quejas: la evaluación, y en especial la calificación numérica, puede verse afectada por factores no controlados como el cansancio del profesor, los prejuicios sobre las capacidades del estudiantes, el orden en que se produce la corrección y algunos otros más (Wolfe et al, 2016).

Los profesores más conscientes elaboran estrategias para combatir estos problemas –corregir por preguntas, en vez de por exámenes completos, o dedicar un tiempo limitado para evitar el cansando, revisar dos veces la evaluación o emplear rúbricas de corrección detalladas, por mencionar sólo algunas. En los

casos extremos, la conciencia de este problema puede llevar a sustituir este tipo de evaluación "subjetiva" por pruebas "objetivas" de tipo test.

Aunque se trata de un problema de difícil solución, no estaría de más que los profesores empezáramos a asumir que esta subjetividad existe, y a tratar primero de medirla, segundo de establecer sus determinantes y en tercer lugar establecer mecanismos para corregirla. Aunque nos cueste admitirlo de cara a nuestros estudiantes, la subjetividad existe, y no sólo en este gremio: cuenta D. Kahneman (2013: 64) de un estudio sobre ocho jueces israelíes que examinaban solicitudes de libertad condicional en el que el principal determinante del resultado (concesión o denegación) resultó ser el cansancio y el hambre de los sujetos, según se alargaba el trabajo y se alejaba el momento de la última ingesta de alimento. Lo nuestro, desde luego, no es tan grave.

Pero nunca viene mal tratar de estudiar un problema. Desde luego, todos somos conscientes de la variabilidad de las notas entre distintos profesores ("huesos" vs. "benditos"), incluso cuando existe un examen común sobre la base de un programa común. Incapaces de encontrarle una solución que no sea establecer exámenes de tipo test, preferimos ignorar el problema, pero éste existe.

La UNED se presta particularmente, por el número de exámenes corregidos, a obtener datos al respecto. La abundancia de exámenes obliga a que varios profesores corrijan las mismas preguntas –procediendo a un reparto por provincias, que cambia de una convocatoria a otra—y permite detectar la existencia de una notable variabilidad en las notas, pese a que hace años que trabajamos con rúbricas o plantillas de corrección, que indican con gran detalle los aspectos a valorar. Pero la interpretación de estas indicaciones parece distar de ser homogénea.

Tabla 1. Diferencial de notas promedio según correctores
(Historia Económica, ADE, Junio 2014)

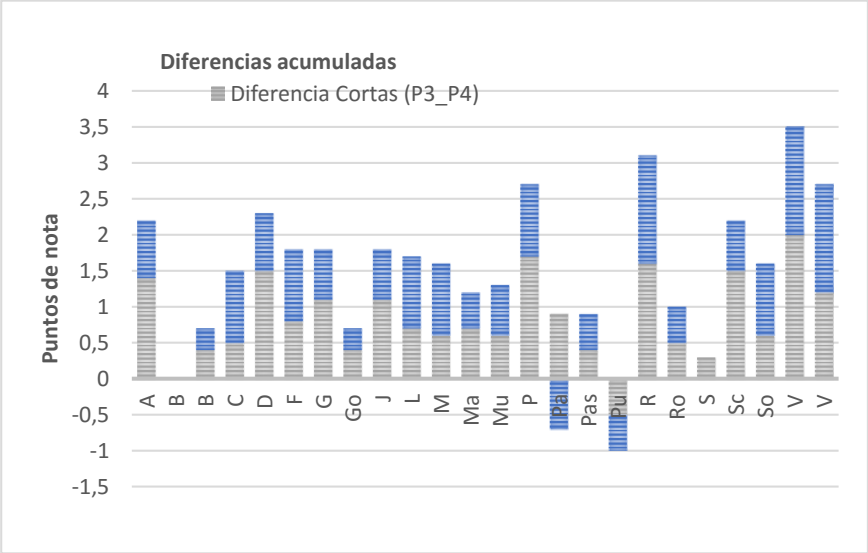
	2011/12				**2012/13**				**2013/14**				**2014/15**			
P=profesor	P1	P2	P3	Todos	P1	P2	P4	Todos	P1	P2	P4	Todos	P1	P2	P4	Todos
Pcortas (/5)	1,7	1,9	2,0	1,9	1,0	1,7	1,2	1,3	2	2	1,8	1,9	1,2	1,4	1	1,2
Comentario(/3)	0,8	1,2	1,3	1,0	0,8	1,3	0,9	1,0	0,9	1,1	1	1,0	0,9	0,9	0,9	0,9
Test (/2)	0,93	1,0	1,0	0,97			s.d.		0,94	0,96	0,94	0,94	1,06	0,99	1,07	1,05
exámenes corregidos	269	255	131	655	363	392	316	1071	275	320	361	956	260	269	268	797

Fuente: Datos de calificaciones, 2011/2015

Para facilitar la comparabilidad de los datos, se han sacado del análisis algunos exámenes (procedentes de centros penitenciarios) que habitualmente corrige un mismo profesor, y suelen obtener calificaciones bajas; también se ha utilizado el mismo examen (junio 2ª semana siempre), por entender que podía haber algunas diferencias en la autoselección de los alumnos en las dos semanas de exámenes. Como puede verse, la variabilidad en las notas (para un mismo examen, no de un

año a otros, que podría atribuirse a muchos otros determinantes) es notable, entre 0,5 y 1 punto sobre un valor de 8 para la suma de las dos partes del examen (comentario de un material de prácticas y cinco preguntas cortas). Afecta por separado y en el mismo sentido a los dos tipos de ejercicio y sin embargo no es visible en el tercer componente del examen , las 10 preguntas tipo test, cuyos resultados no varían más allá del rango previsible por puro azar. Esto parece indicar que no se trata de una diferencia achacable a los estudiantes en sí, su formación previa o su reparto geográfico (aunque sabemos que hay cierta variabilidad basada en la provincia de residencia), sino a preferencias e inclinaciones de los profesores, que se muestran en pautas clara y repetidas. Casi sistemáticamente P1 (nombre supuesto) otorga notas un 10% más bajas de la media, mientras que P2 suele puntuar entre un 15-20% por encima de la media, y P3 un poco más incluso. En cambio, P4 pertenece al grupo de los "huesos", aunque ligeramente menos que P1. Sólo los resultados del último curso para el que hay dato (2015/16) parecen mostrar que esas pautas han dejado de actuar, tal vez como resultado de la armonización derivada del uso de rúbricas, o de una actuación semiconsciente de los profesores para hacer converger las calificaciones, a la vista de los debates dentro del equipo sobre la existencia del problema y la necesidad de buscar soluciones. Estas variaciones de nota pueden no parecer grandes, pero a menudo suponen la diferencia entre el aprobado y el suspenso, y son difíciles de justificar en unas circunstancias en las que la variación es meramente fruto de la "suerte" que le toque al estudiante al serle asignado uno u otro corrector.

Figura 1. Diferencias de notas en exámenes corregidos por duplicado (2012)*



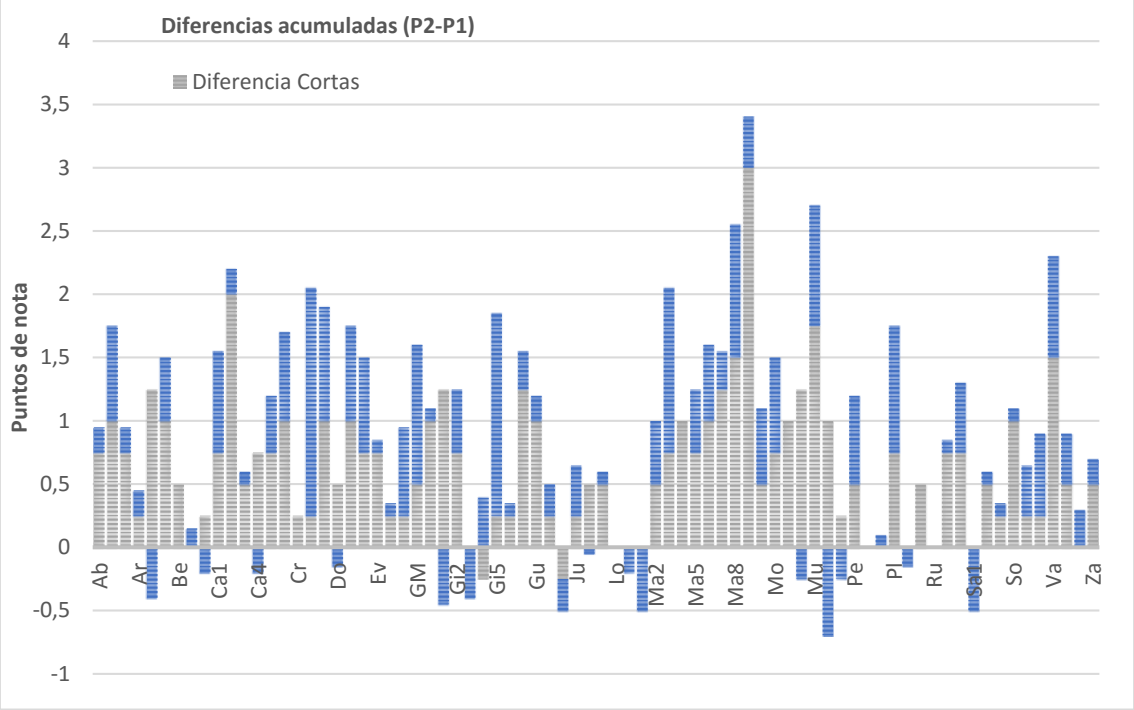
*Datos de 24 exámenes de Historia Económica del centro Barcelona-CUXAM

Las diferencias entre los dos correctores ascienden a 1,5 puntos sobre 8 y de 0,65 puntos sobre 3 en el caso del comentario (de texto, gráfico, mapa o tabla numérica). De nuevo, la pauta se repite, con un profesor hueso (P4 en este caso) frente a otro generoso (P3) que sistemáticamente (con dos excepciones) asigna notas mejores que su colega: de hecho, sólo en un caso hubo coincidencia en la

calificación, y era un examen pésimo (que obtuvo un 0). Como resultado de estas variaciones, en 9 de estos 24 exámenes era la diferencia de un suspenso a un aprobado.

La corrección por duplicado volvió a repetirse (por última vez, esperamos) en junio de 2013, y esta vez afectó a un número mayor de exámenes, y a los otros dos correctores P1 y P2 (figura 2). De nuevo hay diferencias en las notas, menores pero aún sustanciales (0,9 puntos de media sobre 8 totales), que afectaban a 49 de los 75 exámenes y con diferenciales en la calificación de hasta 3,4 puntos!. En 16 de estos casos la diferencia afectaría al aprobado/suspenso del estudiante en cuestión. De nuevo con la pauta habitual: P1 es "hueso" y P2 un "bendito". Pero

Figura 2. Diferenciales de notas en exámenes corregidos por duplicado (junio 2013)*



*Datos de 76 exámenes de Historia Económica del centro Valencia-Alzira (Junio 2013)

La variabilidad detectada persiste (aunque los datos de 2015 de la tabla 1 son esperanzadores), pese a los esfuerzos realizados por el equipo docente para detectarla primero y ponerle coto después: el empleo de una plantilla de corrección consensuada, que desde 2014/15 se convirtió en una rúbrica con puntuaciones asignadas a cada parte, acuerdo sobre los tramos para la calificación (homogeneizándolos en saltos de 0,1 puntos, cuando antes había algunos profesores que afinaban a la décima y otros a 0,25). Con todo, aún no podemos decir que se haya eliminado, aunque sí parece estarse reduciendo (ver tabla 1). En algún momento en todo caso el equipo docente deberá tomar una decisión sobre qué márgenes de variabilidad son aceptables. Los que revelan los ejemplo de doble corrección, desde luego, no lo son.

Como quiera que sea, esta variabilidad podría justificar las alegaciones de "subjetividad" o "injusticia" por parte de los estudiantes, restando legitimidad a las calificaciones en sí mismas y a los profesores que las producen, y alimentando resquemores y desconfianzas que interfieren con la docencia.

Aunque tales situaciones son del todo comunes en la docencia presencial, donde tanto las profesoras "hueso" como los profesores "madres" conviven, y ofrecen a los estudiantes la posibilidad de ajustar sus expectativas cambiándose de grupo, no por ello dejan de ser indeseables. La diversidad en los estilos de enseñanza no debería estar reñida con el principio de que un mismo título debe reflejar un nivel de conocimientos (aptitudes, etc.) iguales en alumnos con calificaciones iguales. Con márgenes de error, obviamente, pero sin grandes disparidades. Esa idea está detrás de los "exámenes de cátedra" que cada vez parecen difundirse más, y tal vez debería llevar a una distribución de su corrección a profesores distintos del que tiene al estudiantes en su grupo.

En una universidad masiva como la UNED, donde no existen grupos, la variabilidad es más difícil de defender. De ahí en parte la popularidad entre los estudiantes de los exámenes tipo-test, y de ahí el atractivo que para quienes somos críticos con el tipo test tiene una herramienta de corrección automática, que puede servir no sólo para liberarnos de una de las tareas más ingratas de nuestro oficio (la corrección) sino que además puede hacerlo de forma ventajosa en términos de estabilidad y fiabilidad de las calificaciones. De modo objetivo. E incluso en la precisión y amplitud del *feedback* (personalizado, aunque no personal) que ofrecen al alumno. En todo caso, siempre queda la opción de utilizarlo como una herramienta que proporcione un borrador, dejando al corrector humano el ajuste fino y personal, en un uso semejante al que podemos hacer del software de traducción.

Por eso, cuando a los profesores firmantes se nos presentaron dos colegas de psicología del aprendizaje de la UNED con la oferta de poner a prueba un sistema de corrección automática de preguntas cortas que parecía funcionar razonablemente no dudamos en apuntarnos. Lo que presentamos aquí son los resultados de los dos primeros años de pruebas con alumnos de la asignatura de Historia Económica, de primer curso del grado de ADE.

2. G-Rubric-Gallito API: un sistema automático de evaluación de respuestas libres a preguntas cortas

Existe entre los profesores la convicción, no sabemos si respaldada por evidencia sólida, de que las respuestas libres de desarrollo, o incluso las respuestas cortas constituyen una de las herramientas mejores para evaluar los resultados de aprendizaje y promover un estilo de aprendizaje profundo en diversos niveles educativos, y desde luego en la universidad. Con todo, el número de estudiantes, la necesidad de diversificar la oferta docente y el peso que ha cobrado la evaluación continua han generado una carga de trabajo que hace muy difícil mantener un examen basado en respuestas de tipo libre. Aun cuando no se sucumba a la tentación del test de elección múltiple y se mantenga un examen

final que exija al estudiante redactar, el profesorado no está en condiciones de ofrecer/corregir muchos ejercicios de este tipo a lo largo del curso, ni mucho menos proporcionar a los estudiantes un *feedback* personal y detallado en forma de comentarios y sugerencias concretas a su ejercicio.

Los estudiantes, por su parte, consideran que los ejercicios de texto libre son pasto de evaluación subjetiva, cuando no directamente caprichosa, algo que consideran—con cierta razón-- "injusto" (Valenti, Neri & Cucchiarelli, 2003). Al no obtener un *feedback* adecuado o instrucciones lo suficientemente precisas, buena parte del potencial de este tipo de ejercicios se dilapida.

El desarrollo de herramientas para la evaluación automática de ejercicios de texto libre ha sido durante década una posibilidad muy remota. Se podía, claro, corregir con rapidez ejercicios de elección múltiple, verdadero/falso, rellenado de huecos y un largo etcétera, pero no preguntar: "¿A qué llamamos revolución industrial?" y que un ordenador evaluara, con suficiente precisión, la respuesta, a falta de los instrumentos conceptuales e informáticos adecuados. Esa situación ha cambiado radicalmente con los desarrollos de la inteligencia artificial en las últimas décadas. Hoy ya tenemos herramientas que permiten a los profesores (y a otros sectores de actividad, con otros fines), evaluar respuestas a preguntas como éstas, pudiendo por tanto liberar tiempo para actividades de mayor valor añadido, sin dejar por ello de proporcionar a los alumnos una evaluación más objetiva que la humana e incluso un *feedback* más preciso que los profesores normalmente están en condiciones de ofrecer. Con un par de ventajas añadidas: la inmediatez en la respuesta y la posibilidad de reiterar el proceso cuantas veces se quiera, a costes mínimos.

Los sistemas de evaluación automática de texto libre son extraordinariamente complejos, y sus aplicaciones desbordan con mucho el ámbito de la enseñanza. Estos sistemas se han basado en distintos enfoques, entre los que destaca el Análisis Semántico Latente (*Latent Semantic Analysis* o LSA). En la última década la investigación está viviendo un boom, con particular hincapié en su aplicación a la enseñanza, aunque sorprendentemente más en entornos reducidos (clases de pocos estudiantes) que en instituciones a distancia con centenares o miles de estudiantes por asignatura o en otras enseñanzas masivas como los *MOOC*s (Jorge-Botana et al, 2015).

2.1 ¿Qué es el LSA?

El LSA se basa en el concepto modelos de espacio vectorial. Lo que se hace es emplear el álgebra lineal para distribuir unidades léxicas en un espacio n-dimensional. "En términos generales, el LSA es un conjunto de diferentes procesos a través de los cuales una colección de textos (manuales, obras de referencia, etc.), normalmente denominada corpus, se transforma en un espacio semántico. Tras procesar el corpus, este se expresa en una matriz en la que se incluyen tanto términos individuales como párrafos. A continuación se le aplica un nuevo proceso a la matriz, incorporándole una función de ponderación, normalmente de tipo log-entropía, al objeto de compensar la asimetría en la frecuencia de las palabras" (Jorge-Botana et al, 2015).

El peso entrópico global que se asigna a cada término nos indica el grado de focalización que presenta respecto a un área temática; lo que es lo mismo, en qué medida el término es específico. Así, un término como lanzadera volante, por ejemplo tiene un peso muy elevado, lo que indica que sólo se usa en determinados contextos, en conexión con otros términos igualmente específicos. Por el contrario, un peso global bajo indica un alto grado de generalidad (como trabajador, productividad o renta, por ejemplo). Este índice permite predecir, una vez que aparece el término, el asunto del que trata el discurso (la respuesta analizar, en nuestro caso). Las ventajas del LSA nacen, no obstante, de la aplicación a esta matriz una técnica de reducción de dimensiones (denominada Descomposición en Valores Singulares (o *Singular Value Decomposition*, SVD), que permite identifica cada término o párrafo con tan sólo unas pocas dimensiones (unas 300), omitiendo el resto, que constituyen el "ruido" presente en el uso común del lenguaje.

Hasta aquí, en grandes trazos, cómo nos describen nuestro sistema basado en LSA sus diseñadores. Para los profesores de historia económica involucrados en la prueba, los rasgos más importantes del sistema radican en la capacidad de ofrecer al estudiante al menos tres tipos distintos de *feedback* a cada una de sus respuestas: una nota numérica para el contenido, otra adicional para la calidad de la expresión escrita y un *feedback* gráfico detallado que representa la cercanía de la respuesta al óptimo en relación a una serie de "ejes conceptuales" que representan una serie de vectores definidos cada uno por alunas serie de palabras-clave (de 5 a 10).

Para alimentar el sistema, los profesores proporcionan dos tipos distintos de inputs.

a) La materia prima para el corpus, con una serie de textos generales. Estos se despojan de todo el material lingüístico "sin significado" (incluyendo pronombres, números, nexos, etc.)— antes de conformarlos en un espacio semántico. Como hemos dicho, este espacio semántico "almacena" también la posición relativa de las palabras en los textos empleados, asignándoles un valor numérico (denominado peso entrópico global) que nos indica el grado de focalización del término (en qué medida es específico o revelador de un determinada región semántica) y a la vez indica su cercanía a otros términos.

Todos estos procesos se realizan mediante Gallito Studio, y el espacio resultante se sube a GallitoAPI, el cuál es un al API (*Application Programming Interface*) que permite que la funcionalidad que puede ser llevada a cabo con este espacio funcione como Servicio Web y pueda ser invocada mediante protocolos estándar (se envían los textos por http y se recibe el análisis mediante el mismo protocolo).

Esta tecnología ha sido desarrollada por investigadores del departamento de Psicología Evolutiva y de la Educación.
La interfaz web para la evaluación de texto libre recibe el nombre de G_Rubric, nombre con el que solemos referirnos al conjunto, aunque es importante tener presente que la creación y gestión del espacio multivectorial que constituye el

corazón del sistema, se realiza a través de Gallito Studio y GallitoAPI (G-Rubric envía los textos a galitoAPI y recibe y presenta los resultados).

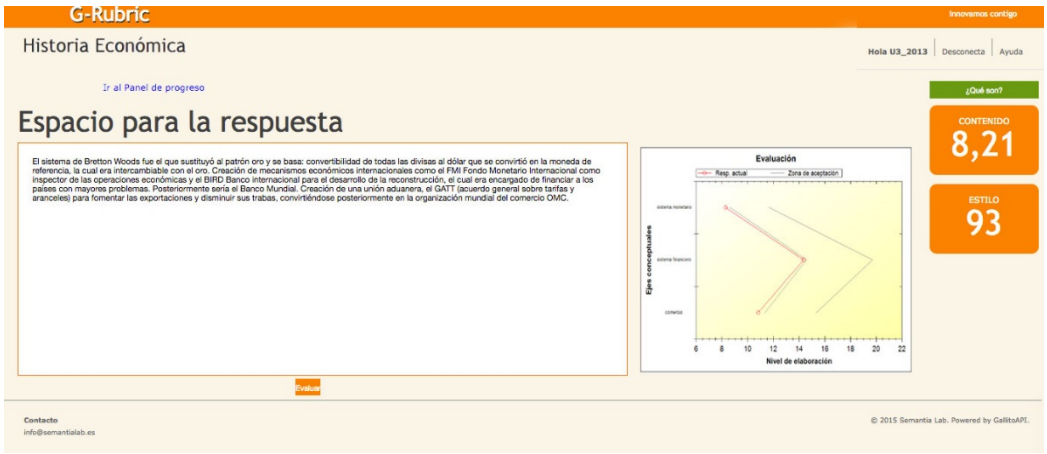
En nuestros caso, el corpus se basó en seis manuales distintos de historia económica mundial, todos ellos escritos en castellano (no traducidos) y publicados en los últimos 20 años (Bernardos, Hernández & Santamaría, 2014; Comín, 2011; Comín, Hernández & Llopis, 2005; Mateos Royo,2014; Palafox, 2014; Simón,1996).

b) La preguntas y respuesta concretas, a las que en adelante nos referiremos como 'objetos'). Se trata de preguntas de texto libre que requieren respuestas cortas (70-200 palabras) suficientemente predecibles. Los estudiantes escriben la respuesta a su gusto, aunque se les indica un rango aproximado de palabras, dado que el sistema penaliza el lenguaje ampuloso. También se les indica qué partes del texto pueden repasar para responder la pregunta.

Cada pregunta va acompañada de un canon de respuesta (o "texto de oro"), con el que se compararán las respuestas de los estudiantes, y una serie de ejes conceptuales (entre 3 y 5, según el objeto) compuestos por varias palabras-clave (entre 5 y 10 por eje) que deben delinear distintas regiones del campo semántico que debe cubrir la respuesta. El diseño del texto de oro y los ejes es uno de los momentos críticos del proceso, empleamos respuestas reales de alumnos a preguntas de exámenes pasados para para testarlos y afinar la fiabilidad tanto de las notas numéricas como del feedback gráfico. El proceso requirió en todos los casos varias iteraciones del proceso hasta que llegamos a niveles que consideramos aceptables para pasar a ser probados por los estudiantes.

Para ello, pusimos en marcha un servidor con un interfaz de web de G-Rubric (figura 3) que permite a los usuarios registrados seleccionar cada una de las preguntas y enviar sus respuestas, que reciben a casi de inmediato el *feedback* previsto (notas y gráfica).

Figura 3. Interfaz web G_Rubric



3. Dos pruebas de G-Rubric con estudiantes de historia económica

Es importante dejar claro que las pruebas se han realizado en un contexto de *evaluación formativa*, no *sumativa*, en la jerga de los pedagogos. Es decir, aspiraban a facilitar el aprendizaje y no a calificar ejercicios de cara a la nota final. El uso de G-Rubric permite ambas posibilidades, aunque sus creadores se inclinan más por su uso en evaluación formativa. Para esta presenta además dos ventajas importantes: permite repetir cuantas veces se determine la actividad y proporciona un *feedback* inmediato. En parte por eso, y en parte por limitar los experimentos que afecten a las notas a la fase de la gaseosa, preferimos de momento no aplicarlos a la *evalución sumativa*. Aunque es evidente que es una de las posibilidades más atractivas del sistema, en la medida en que permite ahorrar trabajo rutinario y proporcionar una solución a los problemas de "subjetividad" y variabilidad puestos de relieve en la primera parte de esta comunicación.

3.1 Primera prueba (mayo de 2015)

Ambas pruebas fueron llevadas a cabo con estudiantes de Historia Económica, una asignatura de 1º del grado de ADE, 2º cuatrimestre, pero con enfoques muy distintos. En la primera prueba (2015), los objetivos se dirigían en dos direcciones: determinar la eficacia de G-Rubric –y más allá, de los distintos tipos de *feedback* ofrecido-- para promover el aprendizaje y establecer la fiabilidad otorgada por los usuarios a la calificación ofrecida por la aplicación. Se buscó por tanto un diseño experimental enfocado a estos objetivos, con dos hipótesis:
H1: El aprendizaje mejoraría en función de la "riqueza" del *feedback* ofrecido.
H2: La mejora del aprendizaje iría en proporción al número de intentos.

Para ello, se buscaron voluntarios entre el conjunto de los estudiantes, con un pequeño incentivo (0,25 puntos de nota) y los 132 voluntarios se distribuyeron aleatoriamente en tres grupos, a los que se asignaron tres tipos de tratamiento distinto (tabla 2), de los que el grupo 3, del que no se esperaba ninguna mejora en término de aprendizaje derivado del uso de G-Rubric, funcionaría como grupo de control.

La prueba exigía de los participantes responder, a través de G-Rubric, a cinco preguntas de respuesta corta, muy similares a las que figuran en el examen, relativas a (1) las características principales de los regímenes demográficos, (2) las consecuencias de la revolución neolítica, (3) los rasgos de las economías agrarias medievales en Europa, (4) el mercantilismo y (5) el comercio triangular en el Atlántico. Para cada pregunta se les indicaba en qué parte del material de la asignatura podían buscar la respuesta, se les daba un rango (+/-20) de palabras que debía tener su respuesta, instrucciones de manejo de la aplicación e indicaciones –diferentes según los grupos— sobre cómo interpretar el *feedback* que esta ofrecía. Los dos primero grupos tenían hasta 6 intentos para cada pregunta, en los que debían emplear el *feedback* recibido y consultar el material (aunque no compartir respuestas entre ellos) para mejorar su respuesta, aunque no se les obligaba a agotarlos, sino que podían detenerse cuando consideraran

satisfechos. La prueba se realizó en mayo de 2015, se explica con más detalle en Hernández et al (2015).

Tabla 2. Ensayo 2015. Tratamientos experimentales

	Tratamiento	Nº intentos /actividad	Número sujetos	De ellos, mujeres	De ellos, repetidores
1	*Feedback* "rico": gráfico y numérico.	6	44	30	9
2	*Feedback* "pobre": sólo numérico	6	44	29	8
3	*Feedback* no utilizable: un único intento	1	44	30	13

Como medida del "aprendizaje" se tomó la nota obtenida una única actividad final, con un único intento para todos los participantes. También se emplearon los diferencial de calificación entre el primer/último intento, y entre el mejor/peor intento. Adicionalmente, se trató de medir por otra vía, considerando las notas medias de los distintos grupos de voluntarios en las repuestas a las preguntas cortas del examen.
Como medida de la "fiabilidad" percibida se empleó una encuesta posterior de respuesta anónima.

Por desgracia, los resultados resultaron poco concluyentes, al menos en términos de aprendizaje (tabla3). Es cierto que la aplicación consiguió cierta adhesión de los participantes, que en muchos casos agotaron los intentos disponibles. También que, en general (pero no siempre, ver pregunta 3), la nota media en todas las actividades era sistemáticamente superior para G1 (*feedback* rico) que para los otros dos grupos. También, como norma, la nota de G1 y G2 era superior a la de G3, indicando que la reiteración de intentos ayudaba al aprendizaje. También puede adivinarse el efecto del *feedback* rico en la diferencia promedio max/min por usuario (salvo en la actividad 1) que eran creciente, especialmente en G2, podría indicar que la repetición del ejercicio reforzaba el aprendizaje. Pero no son diferencias muy concluyentes. En sentido contrario, en cambio, apuntan el principal indicador de aprendizaje previsto, la calificación obtenida en la prueba final de control, que no muestra diferencias significativas en ninguno entre las tres condiciones experimentales (ya sea con *feedback* rico, pobre o con *feedback* no utilizable).

Tabla 3. Ensayo 2015.Indicadores de aprendizaje

		Nota media G-Rubric (/10)			Diferencia max-min		
	Pregunta	G1	G2	G3	G1	G2	G3
1	Regímenes demográficos	6,9	6,5	6,4	0,52	0,69	0
2	Consecuencias de la revolución neolítica	6,5	5,9	5,6	1,06	0,95	0
3	Economías agrarias medievales Europa	6,2	7,4	5,5	1,10	0,78	0
4	Mercantilismo	7,7	7,5	6,6	1,95	1,15	0
5	(Final) Comercio triangular	6,2	6,3	6,1	0	0	0

La comparación con los resultados de examen tampoco arrojó resultados muy concluyentes (tabla 4). Para empezar, en contra de la hipótesis inicial, los resultados de G2 (*feedback* pobre) resultaron de media mejores que los de G1 (*feedback* rico), que no diferían de los de G3 (control). Por otro lado, . aunque los tres grupos obtuvieron en las preguntas cortas notas superiores a las medias del alumnado (entre un 25-40% más alta), este efecto se producía también en las notas de la parte de comentario y de test, que no se trabajaban específicamente vía G-Rubric. De hecho, en general todo el grupo de voluntarios (independientemente de su condición) obtuvo casi 1,5 puntos de nota más que la media. Es por tanto probable que el determinante de estas diferencias no sea la participación en G-Rubric sino un sesgo de autoselección que hace que los voluntarios reunieran de antemano otras características (seguimiento más activo de la asignatura, mayor motivación u otras) que explicarían, a la vez, su participación en la prueba y sus calificaciones superiores a la media.

Tabla 4. Ensayo 2015.Notas medias en las distintas partes del examen, de los voluntarios G-Rubric y el conjunto de alumnos presentados a examen

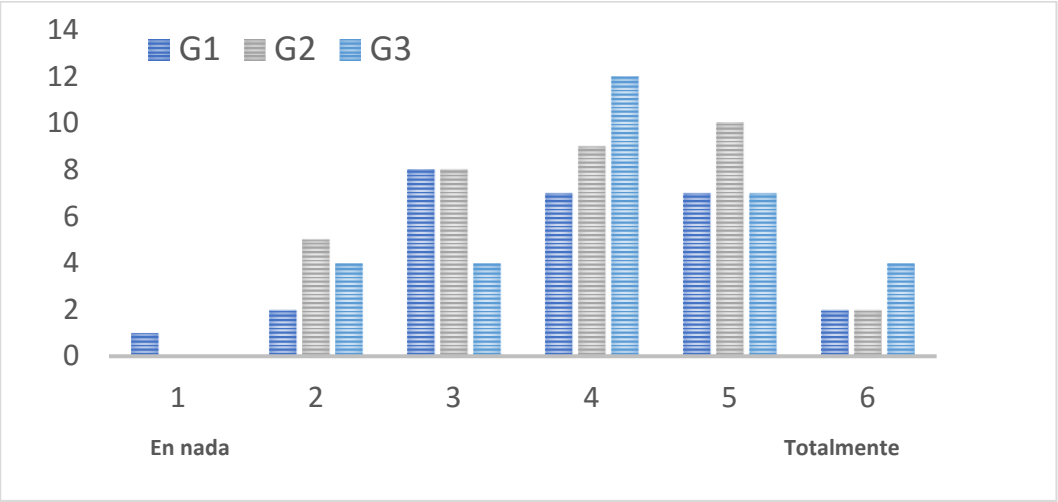
Nota medias según partes del examen	G1	G2	G3	Presentados a examen en junio
Preguntas cortas (/5)	1,9	2,4	1,9	1,4
Comentario (/3)	1,36	1,37	1,22	1,04
Nota media test (/2)	1,20	1,41	1,28	1,09
Total examen(/10)	4,43	4,85	4,42	3,81

En suma, nuestra prueba no consiguió más que indicios poco concluyentes en relación con nuestras hipótesis (H1: El aprendizaje mejoraría en función de la "riqueza" del *feedback* ofrecido;H2: La mejora del aprendizaje iría en proporción al número de intentos).

Tras analizar los datos, los autores consideramos que debía haber problemas no previstos en el diseño experimental y/o el control de su desarrollo, como por ejemplo que el plazo en que se hicieron las pruebas resultó demasiado corto, que los objetos de G-Rubric no estaban bien construidos, que el número de pruebas/intentos era demasiado escaso para promover el aprendizaje y que deberíamos tratar de controlar el sesgo de autoselección mediante un procedimiento distinto de selección de sujetos.

En cambio, los resultados en términos de satisfacción (figura 3) –medida en encuesta-- con la aplicación, de cara a determinar la fiabilidad percibida, fueron bastante más en línea con lo previsto, y bastante satisfactorios en general, aunque de nuevo no mostraban grandes diferencias entre grupos.

Figura 4. Indicador satisfacción1. Exactitud percibida de las notas de G-Rubric



Nota**:** Repuestas a la pregunta "¿En qué grado está conforme con la CALIFICACIÓN recibida en el ÚLTIMO INTENTO de respuesta?" (G1 y G2) y "¿En qué grado está conforme con las CALIFICACIONES recibidas? (G3)

Los resultados muestran una cierta conformidad (entre 3,6 y 4,1 sobre una escala de 6), curiosamente mayor en los voluntarios que hicieron menos intentos (G3).

Este hecho, sumado a nuestra propia convicción de la posibilidades de esta herramienta de cara al futuro nos animaron a replantear la experiencia, y repetir la prueba en el siguiente curso, aunque con parámetros distintos.

3.2 Segunda prueba (abril/mayo de 2016)

El planteamiento de la segunda prueba yo no fue testar la capacidad de aprendizaje en condiciones de control experimental, sino fundamentalmente mejorar el diseño de los objetos (preguntas) para tratar de potenciar el aprendizaje y la satisfacción de los usuarios. Esta, de nuevo, se ha medido vía encuesta tras la prueba, mientras que el aprendizaje se medirá exclusivamente por la puntuación de los voluntarios en las preguntas cortas del examen de junio. En cuanto a los cambios introducidos, se probó con más tiempo y cuidado cada objeto, se aumentó su número a siete, se redujo el número de intentos a tres y se mejoró el *timing* de las pruebas (dando más plazo y haciéndolo más flexible). Al objeto de embarcar a más voluntarios, se les ofreció un incentivo de un punto de nota (algo ficticio, pues en realidad se les detraía de la parte de la evaluación continua), que además dependería no de sus resultados calificaciones), sino de su grado de actividad (número de intentos), como modo de minimizar los posibles fraudes, imposibles de controlar en una situación como la de al UNED, con actividades realizadas *on line*.

En cambio, al no tener que establecer grupos con tratamientos distintos, pudimos difundir la información general, y resolver dudas de funcionamiento empleando los foros del curso virtual, lo que facilitó mucho la tarea con respecto al curso anterior.

Por desgracia, la prueba está en marcha en el momento de redactar estas páginas (se cierra el 31 de mayo, cuando nuestro plazo de entrega acaba el 27), por lo que los datos que podemos presentar, tanto de participación como de encuesta son no sólo provisionales sino incompletos. Se trata de un anticipo obtenido en la fecha más tardía posible (24 de mayo, 2016). Por otro lado, como los exámenes de junio se celebran también después del encuentro, no podemos comparar los resultados de participación con las notas de examen.

Las actividades (objetos) en esta prueba fueron los siguientes. Las cifras de son provisionales; hasta el 24 de mayo había 99 estudiantes inscritos, aunque aún podrían inscribirse alguno más hasta la fecha de cierre, pero sobre todos tienen 7 días más para completar las actividades.

Tabla 5. Ensayo 2016.Indicadores básicos

		Num Resps.	Notas promedio		Notas individuales		
	Objeto	# resp	nota max	nota min	max	min (>0)	texto de oro
1	Mercantilismo	248	7,5	6,1	9,67	0,7	9,67
2	Comercio triangular	216	5,7	4,4	7,5	0,6	9,19
3	Carbón mineral en la Rev Industrial	177	6,3	5,0	9,57	0,8	9,57
4	Ventajas rezagados (Gerschenkron)	162	5,9	4,7	9,06	1,2	9,06
5	Segunda Revolución Industrial	158	7,1	5,4	9,21	0,6	9,29
6	Consecuencias económicas Primera GM	135	6,5	4,9	8,17	2,7	9,38
7	Tres pilares de Bretton Woods	128	4,1	5,7	9,33	0,7	9,33

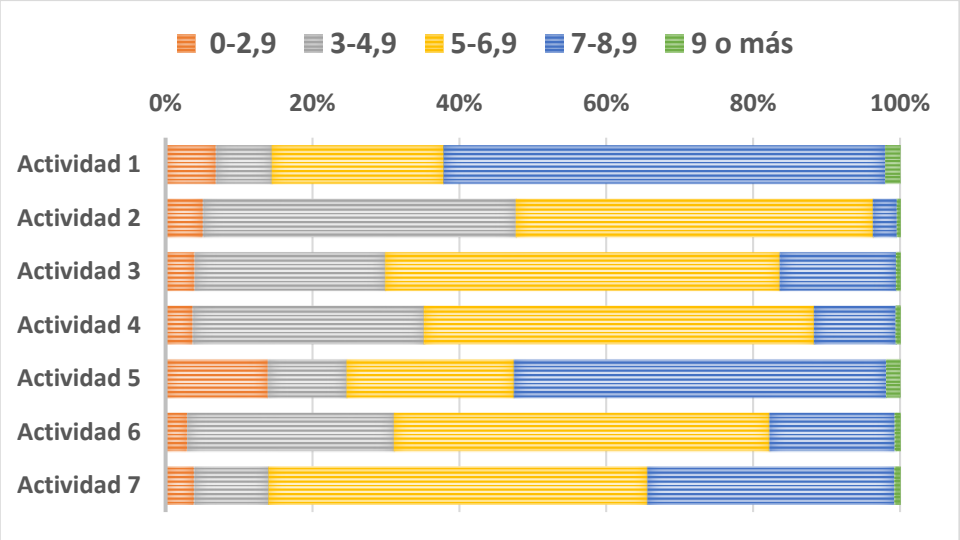
Dicho esto, tenemos algunos datos de interés.

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 ultimo recibían *feedback*, pero ya no había posibilidad de rectificar). 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). Las notas medias son coherentes con las producidas en el ensayo de 2015 (ver tabla 3), aunque debe tenerse en cuenta que entonces tanto el G1 como el G2 tenían hasta 6 intentos, frente a 3 máximo en 2016. Quizá el dato más importante lo dan las notas máximas absolutas, que en todas las actividades superan el 7,5, y en 5 de 7 superan el 9 (hay que aclarar, por otro lado, que la puntuación obtenida con la llamad0 “texto de oro” no es nunca un 10, debido a una función de ajuste del sistema; ver tabla 5). Las notas mínimas son normalmente muy bajas, pero puede deberse a errores de uso o mala comprensión de lo que se esperaba del usuario. Pero a cambio el número de respuestas que superan el 9 son un porcentaje considerable respecto a las respuestas totales.

En suma, la aplicación funciona como herramienta de evaluación y de forma razonablemente satisfactoria, como lo indican las respuestas de los "usuarios" en encuesta.

Adicionalmente, la distribución de notas por objetos (figura 5) nos permite testar el grado de dificultad/calidad de cada uno, lo que puede resultar muy útil para la programación de las actividades. De hecho, estos datos podrían explicar en parte por qué los datos del ensayo de 2015 no fueron muy concluyentes: la prueba final, a un único intento, era la pregunta sobre el comercio triangular (actividad 2 en 2016), cuyos resultados (notas máximas, medias y distribución) revelan que resulta particularmente difícil (o está mal diseñada como objeto).

Figura 5. Distribución de las calificaciones por objeto/actividad



El análisis del aprendizaje, tomando como indicador diferencia entre el mejor y el peor intento en cada actividad, 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 6)

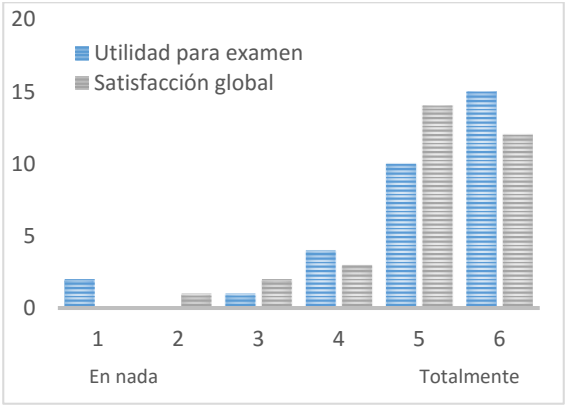
Tabla 6. Ensayo 2016. Aprendizaje. Diferencia entre el mejor y el peor intento, en término absolutos (puntos de nota) y relativos (% de mejora sobre peor respuesta) . Promedios.

	Identificador de actividad							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**Todas**
Absoluta (puntos de nota)	1, 5	1,3	1,3	1,2	1,7	1,6	1,3	1,4
Relativa (%)	41,9	34,5	39,2	31,1	67,2	40,7	28,8	40,4

Dado que se trabajamos con datos provisionales, no merece la pena avanzar mucho más en el análisis, aunque puede verse que la muestra ya apunta direcciones interesantes, a falta de pasar la verdadera prueba final que es calibrar el aprendizaje con los resultados de examen real. Queda por fin ver los resultados de las encuestas.

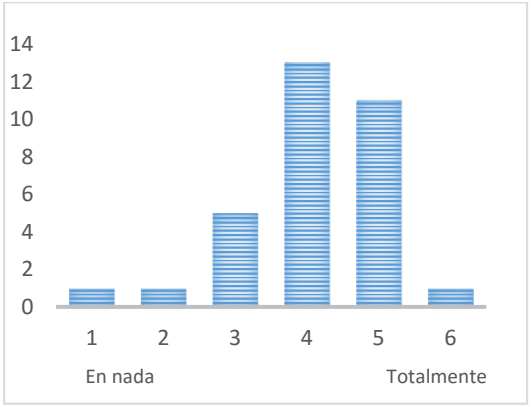
A este respecto, solo podemos anticipar los datos de encuesta, con 32 respuestas (a 27 de mayo) sobre 99 participantes en el ensayo. Como en 2015, los resultados en términos de utilidad percibida/satisfacción con la aplicación (figura 6) y conformidad con la nota obtenida en el ultimo intento (figura 7)–quizá habría que preguntar por el mejor intento, que no siempre es el ultimo-- revelan que los participantes en general entienden que la evaluación automática funciona razonablemente bien y les ayuda a estudiar. Cabe sospechar, además, que si les pasáramos una pregunta semejante en relación con sus notas en examen obtendríamos índices de conformidad quizá menores. Un dato para meditar de cara al uso de G-Rubric en evaluación sumativa.

Figura 6. Ensayo 2016. Utilidad y satisfacción de G-Rubric



Nota: Repuestas a la preguntas "¿En qué grado cree que las actividades de evaluación automática como éstas le servirán para mejorar su preparación de cara al examen? (Utilidad) y "¿Cómo califica globalmente esta experiencia en relación a su proceso de aprendizaje en esta asignatura?" (Satisfacción).

Figura 7. Exactitud percibida de las notas de G-Rubric



Nota: Repuestas a la pregunta "¿En qué grado está conforme con la CALIFICACIÓN recibida en el ÚLTIMO INTENTO de respuesta?"

Además de las respuestas, resultan especialmente alentadores los comentarios libres de los estudiantes, que revelan que al menos para algunos G-Rubric se les ofreció como un desafío mental que tenía mucho de juego:" una gran herramienta, muy práctica, en cierto modo divertida, un poco "juego" y aun consultando apuntes después del primer intento hace sintetizar el tema y repasarlo". Las posibilidades de "gamificar" estas actividades *on line* son sin duda una vía interesante para el futuro (Werbach y Hunter, 2012).

Conclusiones

Pese a los resultados del primer ensayo con G-Rubric, que distaron de ser concluyentes, estamos convencidos de que hubo algún problemas en el diseño o control del experimento que interfirieron en los resultados. De ahí el cambio de enfoque dado al segundo ensayo, aunque no descartamos volver a tratar de establecer la utilidad de la herramienta en condiciones controladas.

En cuanto al análisis de la evaluación humana y automática, hay que repetir que el ensayo de 2016 está en curso. Con todo, este trabajo ofrece algunas conclusiones que pueden considerarse sólidas.

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.

Por último, el desarrollo de los contenidos y la explotación de los resultados de este tipo de sistemas son especialmente adecuados para el trabajo colaborativo: desarrollar cada objeto requiere cierto tiempo, al igual que alimentar el corpus o familiarizarse con la lógica del sistema. Sin embargo, una vez creado puede dimensionarse para un número muy elevado de usuarios, generando un notable incentivo para la colaboración. La elaboración colectiva de objetos, el proceso de testado con estudiantes y por pares, la realización de ensayos y análisis de datos y la utilización habitual por parte de los estudiantes desde cualquier lugar que tenga acceso a internet permiten y alientan esa colaboración. Por nuestra parte, el equipo de desarrollo de G-Rubric y los profesores de historia económica de la UNED os ofrecemos embarcaros en este proyecto, del que creemos que todos tenemos mucho que ganar.

Agradecimientos

Los autores desean agradecer la colaboración de los otros profesores del equipo docente de Historia Económica de la UNED, José U. Bernardos y Rafael Barquín, que como víctimas o cómplices son también protagonistas (¿como profesores P1,P2,P3 o P4?) de este trabajo. Y también especialmente a nuestros compañeros del departamento de Psicología Evolutiva y de la Educación en la UNED, Chema Luzón y Guillermo de Jorge, padres G-Rubric, por invitarnos a acompañarles en este viaje.

Referencias bibliográficas

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.

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

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).

Yousef, A.M.F., 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.