marcoELE
REVISTA DE DIDÁCTICA ESPAÑOL COMO LENGUA EXTRANJERA

RUIZ HERRERO, JUAN
INSTITUTO CERVANTES DE BEIRUT

ESCRIBIENDO A UNA MÁQUINA: LA EVALUACIÓN AUTOMÁTICA DE TEXTOS A DEBATE

BIODATA

Juan Ruiz Herrero es desde septiembre de 2011 profesor interino en el Instituto Cervantes de Beirut. Licenciado en Filología Árabe por la Universidad Autónoma de Madrid, obtuvo en 2011 el Doctorado de Estudios Árabes e Islámicos por el mismo departamento. Ha desarrollado su labor docente en el campo del ELE durante los últimos diez años entre Egipto y Líbano, más concretamente, en los Institutos Cervantes de Beirut y El Cairo, en la Universidad Saint-Joseph (USJ) de Beirut y en la Universidad de El Cairo. Tiene en su haber un diploma de Experto en E/LE (2010) y un Máster de Lingüística Aplicada a la enseñanza de español como lengua extranjera (2014), ambos por la Universidad Antonio de Nebrija. Actualmente realiza un Máster de Exámenes de idiomas en la Universidad de Lancaster (Reino Unido).

RESUMEN
Los sistemas de evaluación automática de textos, programas diseñados para puntuar ejercicios de producción escrita en exámenes en L1 o L2, se han generalizado en las últimas décadas en el panorama académico estadounidense. La presente contribución se propone plantear el estado de la cuestión, describiendo el funcionamiento de dichas aplicaciones y analizando el constructo que miden. Se presentan, así, las ventajas que supone su introducción en el contexto de exámenes de certificación y del trabajo de aula, contrapuestas a los argumentos que hacen valer sus detractores, con el objetivo de acercar la temática al ámbito hispanoparlante y apuntar una reflexión desde la perspectiva del docente de lenguas extranjeras.
PALABRAS CLAVE: evaluación, textos, exámenes, puntuación, validez

ABSTRACT
Automated Essay Scoring (AES) engines, computer systems designs to rate written exercises in both L1 and L2 tests, have known a remarkable spread through the last decades in the academic reality of the United States. The present article reviews the state of the question, describing the functioning of these devices and analyzing the measured construct. The main advantages related to their introduction will then be presented, in contraposition to the main objections voiced by its detractors, with the aim of sustaining a reflection from the perspective of foreign languages teachers.
KEY WORDS: Automated Essay Scoring (AES), scoring, assesment, texts, validity

1. INTRODUCCIÓN

Una de las evoluciones más notables registradas en las últimas décadas en el ámbito de la tecnología informática aplicada a la evaluación de idiomas la ha constituido la generalización de instrumentos de evaluación automática de textos, esto es, programas informáticos calibrados para "evaluar y puntuar la prosa escrita" (Dikli, 2006: 4). Fundamentalnente asociados a exámenes certificativos, tanto de L1 o L2, su uso ha conocido de forma gradual una aceptación considerable en determinados contextos institucionales de los Estados Unidos, con la consiguiente multiplicación de contribuciones científicas acerca de sus características, su pertinencia, así como sus riesgos potenciales. Se trata, de hecho, de un asunto altamente controvertido, con una plétora de destacadas autoridades académicas opuestas a su introducción. A título de ejemplo, la denominada Conferencia de Composición Universitaria y Comunicación publicó en 2004 una declaración formal en la que se desaprobaba explícitamente el uso de este tipo de programas en cualquier contexto formativo, insistiendo en el hecho de que la palabra escrita, en tanto que acto social, debería siempre dirigirse a audiencias humanas, independientemente de su naturaleza y objetivos (Deane, 2013). La polémica resultante llama la atención por su virulencia, con ocasionales intervenciones marcadas por un inusitado dramatismo en el ámbito académico, fiel reflejo del cisma general dentro del campo de la evaluación de la expresión escrita entre los expertos de medición educacional y psicometría por un lado, y los docentes de composición y didácticas de lenguas extranjeras por otro.

Ante la escasez de referencias en español acerca de este fenómeno, la presente contribución se propone ilustrar el estado del debate académico actual acerca de los sistemas de evaluación automática de textos. Se partirá, así, de una explicación del funcionamiento de este tipo de programas, complementada por una exposición de su estatus actual en el ámbito de los exámenes de idiomas. Acto seguido se planteará una revisión de la investigación centrada en determinar la validez de constructo de las pruebas que utilizan dichas técnicas. Más adelante se referirán las objeciones fundamentales enunciadas contra estos sistemas y por último se proporcionará una perspectiva recapitulativa acerca del debate, desde la intención de contribuir con una posición argumentada.

Una cuestión preliminar adicional. Resulta una obviedad apuntar hasta qué punto el debate científico académico en general y más concreto sobre la didáctica de la expresión escrita tiende a discurrir fundamental sino exclusivamente por medios y canales anglófonos. La temática que centra las presentes páginas ejemplifica particularmente tal desequilibrio al tratar un fenómeno que hasta la fecha se ha aplicado exclusivamente en el contexto geográfico particular de los Estados Unidos. En aras de favorecer la lectura, se asumirá en adelante este marco físico sin mencionarlo explícitamente, pero sí que parece conveniente subrayar que dicha omisión no obedece a una pretensión alguna de pertinencia universal de todo aquello que emana del ámbito universitario estadounidense.

Cerraremos la introducción apuntando que este artículo constituye una adaptación de un trabajo producido en el marco del Máster de Exámenes de idiomas de la Universidad de Lancaster.

2. SISTEMAS DE EVALUACIÓN AUTOMÁTICA: ¿QUÉ, DÓNDE Y POR QUÉ?

Los sistemas desarrollados para puntuar automáticamente textos escritos se basan en la inteligencia artificial, un paradigma de investigación transversal cuyo objetivo fundamental lo constituye la simulación de la inteligencia y el comportamiento humanos a través del sistema electrónico de un ordenador (Williamson, 2004). Una de sus áreas más productivas ha resultado ser la del procesamiento de lenguajes naturales, que se centra en los intentos orientados a analizar textos mediate herramientas informatizadas y entre cuyas aplicaciones destacan los programas aquí debatidos. Si éstos fueron concebidos fundamentalmente para asistir en la puntuación de ejercicios y tareas de respuesta elaborada dentro del contexto de servicios certificativos a gran escala, se han venido presentando más recientemente como útiles de aula aptos para la práctica y el desarrollo de habilidades tanto en L1 como en L2. Su dinámica de funcionamiento es siempre la misma: tras haber procesado un número determinado de copias corregidas por humanos de un ejercicio particular, los sistemas son programados para reconocer toda una serie de rasgos lingüísticos a los que se les asignan diferentes valores. A una operación semejante subyace la asunción de que la calidad de un texto se puede establecer mediante un conjunto de dimensiones discretas, reconocibles y sujetas a una posible medición, para cada una de las cuales se diseña una fórmula matemática (Crossley *et al.*, 2014).

Son tres las razones que suelen esgrimirse para justificar la pertinencia del recurso a estos sistemas. En primer lugar y de forma más destacada, la posibilidad de garantizar altos niveles de fiabilidad- sobre todo interevaluadora. Así, aplicar un procedimiento extremadamente objetivo a cantidades ilimitadas de textos homogeneizaría los criterios de calificación y suprimiría la necesidad de realizar sesiones para formar a examinadores, todo lo cual se traduciría en herramientas evaluativas extremadamente estandarizadas y rentables. En segundo lugar, su introducción convertiría la puntuación en un proceso prácticamente inmediato, reduciendo de forma drástica los plazos establecidos para la entrega de resultados. Por último, desde la perspectiva del rendimiento docente, se sugiere el uso de estos sistemas en el contexto del aula, para la práctica o como útiles de puntuación para aliviar la carga de trabajo de los instructores, que podrían delegar las responsabilidades de asignar nota y proporcionar retroalimentación a una herramienta automática.

Resulta oportuno señalar en este momento que por mucho que estos sistemas se puedan antojar extravagancias futuristas, las herramientas automatizadas de evaluación de textos acumulan ya en realidad décadas de historia. Así, en 1968 Ellis Page y Dieter Paulus crearon una primera aplicación experimental destinada a asistir a profesores desbordados para puntuar composiciones en L1 en ámbitos universitarios. Su herramienta informática podía identificar 30 rasgos lingüísticos discretos de los que se pensaba que presentaban altas correlaciones con la calidad de la expresión escrita, tales como la extensión general del texto, la longitud media de las palabras empleadas, la cantidad y clase de recursos de puntuación o el número de errores ortográficos. Si bien consiguieron demostrar una capacidad relativamente alta para predecir las puntuaciones de los examinadores, las instancias académicas de la época manifestaron una oposición férrea a cualquier aplicación del invento, fustigando con dureza la mera noción de escribir a una audiencia compuesta por ordenadores y las implicaciones comunicativas que algo semejante supondría

(Herrington & Moran, 2001). Posiblemente una reacción de tal virulencia explique la ausencia de cualquier contribución de este orden durante las siguientes tres décadas, periodo, además, que en el ámbito de la composición corresponde a la adopción de metodologías centradas en el proceso. En cualquier caso, los sistemas de evaluación automática regresaron con fuerza en la última década del siglo XX, apoyándose en los avances exponenciales registrados en el ámbito de la tecnología informática. Es entonces cuando comenzaron a imponerse como instrumentos legítimos, al menos en determinados círculos académicos.

Así pues, en 1996 un prototipo de lo que más adelante se conocería como Asesor Inteligente de Textos ("Intelligent Essay Assessor") se utilizó para puntuar trabajos escritos en la Universidad estatal de Nuevo México. Se apoyaba para ello en la indexación semántica latente, un modelo estadístico que establece concordancias a nivel de la palabra y del morfema para comparar la semejanza semántica entre diferentes textos. Su versión en línea fue adquirida en 2004 por la editorial Pearson Education. Otros sistemas similares son IntelliMetric, desarrollado por los laboratorios Vantage y posteriormente adquirido por la organización estadounidense College Board, y el "e-rater", creado por la institución ETS[1] y utilizado en 2002 para puntuar exámenes de admisión GMAT (Haswell, 2006). El panorama actual ilustra una considerable proliferación. Así, Shermis (2014) refiere un estudio público en el que siete proveedores comerciales y un laboratorio de investigación puntuaron con sus respectivas aplicaciones diferentes grupos de exámenes provenientes de contextos distintos para comparar su capacidad de adaptación manteniendo un elevado nivel de correlación con la evaluación humana. El autor describe un contexto en los Estados Unidos en el que sucesivas reformas educativas han reforzado la necesidad de demostrar una competencia sólida de expresión escrita para acceder a la educación superior, con la consiguiente multiplicación de tareas de composición en L1 en distintas materias que requieren de una nota o de cierta retroalimentación. Así las cosas, las plataformas educativas oficiales promueven de forma activa el uso de la evaluación automatizada "para permitir puntuar semejante volumen de trabajo de una forma rápida y rentable" (p. 54). Dikli (2006), por su parte, describe el funcionamiento de cinco sistemas diferentes y sus respectivas aplicaciones, al menos dos de las cuales están programadas para procesar textos en otras lenguas además del inglés.

Destacaremos aquí la herramienta "e-rater", puesto que es la primera que se ha aplicado a exámenes certificativos de lengua extranjera. Desarrollada por un equipo de más de quince expertos en procesamiento de lenguajes naturales, lingüística computacional y ciencia cognitiva, la herramienta desarrollada y patrocinada por el ETS es la que ha sido objeto de mayor atención en lo que se refiere a contribuciones académicas (Wiliamson *et al.*, 2012). En 2005 el "e-rater" sustituyó al segundo examinador tradicionalmente empleado para puntuar una de las dos tareas de expresión escrita dentro de la versión adaptada a Internet del examen certificativo TOEFL (Test of English as a Foreign Language). Si la diferencia entre ambas notas (la del "e-rater" y la del examinador humano) resulta menor de 1,5 puntos en una escala de 5, el resultado final es la media entre ambos valores. Si, por el contrario, la diferencia iguala o supera dicha referencia, el texto se entrega a un segundo examinador experto y la

[1] Tanto el College Board como el ETS (Educational Testing Service) son instituciones educativas estadounidenses sin ánimo de lucro fundamentalmente activas en el ámbito del desarrollo de exámenes.

nota final se calcula como la media resultante[2]. Aquellos textos que el programa identifica como anómalos -demasiado largos, demasiado cortos o con demasiados errores- se entregan exclusivamente a examinadores humanos (Enright & Quinlan, 2010).

3. LA VALIDEZ DE CONSTRUCTO DE LOS SISTEMAS DE EVALUACIÓN AUTOMÁTICA

Puesto que la fiabilidad y la eficiencia constituyen los atributos más notables de los sistemas de evaluación automática de textos, la mayor parte de experimentos llevados a cabo por los equipos de investigación que los respaldan se han centrado en manifestar y demostrar la validez de los exámenes que los emplean. Se trata, así, de establecer una cierta legitmidad que sirva para desmontar las aprensiones suscitadas por el hecho de delegar la evaluación a lectores automáticos e inconscientes. Ya que estas aplicaciones reconocen microcaracterísticas que se consideran representativas de la calidad de expresión escrita, el desafío fundamental reside en demostrar empíricamente que la combinación de todas ellas forma un constructo que refleja de forma coherente y consistente aquello que los examinadores humanos entienden como buenas prácticas de escritura. Para ello, como en cualquier otro experimento relacionado con la validez de un examen, se tratará de evitar lo que Messick (1993) definió como las dos mayores amenazas a la validez de constructo, a saber, la infrarrepresentación de constructo y la varianza irrelevante de constructo.

El enfoque de la investigación cuantitativa experimental en este área se ha venido centrando en demostrar una alta correlación con las calificaciones otorgadas por examinadores expertos, lo que sugeriría la habilidad de predecir calificaciones holísticas tradicionales y, por ello, la de medir el mismo constructo (Attali, 2007; Attali & Powers, 2009). No obstante, como Shermis (2014) observa, una coincidencia de ese tipo no implica necesariamente una validez de constructo sólida, sino más bien la conformidad con un criterio particular cuya pertinencia estaría todavía por probarse, ya que "un modelo predictivo puede predecir satisfactoriamente el comportamiento humano al puntuar, pero por razones ajenas al constructo en cuestión" (p. 74). Se ha asistido, pues, a lo largo de los últimos quince años a todo un abanico de perspectivas y enfoques de investigación relacionados con los programas de evaluación automática.

En primer lugar, estos sistemas se han utilizado como herramientas en estudios que tenían como objetivo atender a cuestiones lingüísticas más amplias. Así, por ejemplo, Crossley *et al.* (2014) se centraron en investigar cómo las microcaracterísticas lingüísticas analizadas por los programas pueden llegar a explicar el dominio en una L2. Se utilizaron dos de ellos en el estudio que, a partir de un corpus de 480 tareas escritas del examen TOEFL, analizaba un total de 189 índices operacionalizados como rasgos evaluables. El modelo de regresión aplicado reveló que el predictor más fiable de calidad de escritura era el número de categorías de palabras diferentes utilizadas, lo que constituiría un indicador de la amplitud del repertorio léxico. Otros ragos que presentaban una correlación elevada incluían el número total de palabras en el texto, la mayor abstracción del vocabulario y la proporción de palabras claves asociadas al tema de cada tarea. Una perspectiva similar se adoptó en el estudio de Crossley y McNamara (2012) acerca de los rasgos de

[2] A menos que una de las tres calificaciones difiera de las otras dos en más de un 1,5 puntos, en cuyo caso es anulada.

sofisticación lingüística en L2, en el cual se aplicó CohMetric a un número de tareas producidas por estudiantes de Hong Kong.

En cuanto a la cuestión de cómo estos rasgos reflejarían la validez de constructo deseada, Crossley *et al.* (2014) examinaron los elementos principales que explican valores divergentes entre textos puntuados por examinadores humanos y automáticos. Hicieron así hincapié en la insuficiencia de los criterios programados para detectar la elección léxica, dando a entender la imposibilidad de estos sistemas para identificar la adecuación del vocabulario. No en vano, tal y como apuntan, un lector automático puede detectar con facilidad la corrección ortográfica de una palabra pero si ésta ha sido utilizada apropiadamente o no en un contexto determinado constituye una dimensión que ninguno de los programas de evaluación automática desarrollados hasta la fecha se encuentra en condiciones de valorar.

Es precisamente la validez de los rasgos encargados de medir las dimensiones vinculadas al contenido del texto la que suscitan mayores reticencias a la hora de reflejar el constructo. Puesto que los sistemas de evaluación automatizada aplican fórmulas matemáticas, pueden reconocer de forma eficiente elementos lingüísticos superficiales pero suscitan escepticismo cuando deben evaluar parámetros textuales que resultan más difíciles de operacionalizar con ecuaciones que, por definición, se revelan insensibles a los matices del significado y la expresión. El constructo del "e-rater", por ejemplo, se estructura en nueve grandes áreas: gramática, uso, convenciones gráficas, estilo, organización, desarrollo, aspectos positivos, complejidad léxica y uso del vocabulario específico del tema, cada uno de los cuales se deslinda en una serie de subrasgos. Si bien la atribución de algunos de ellos al bloque de uso en vez de al de gramática, pongamos por caso, pueden suscitar cierta controversia, no existe mayor duda acerca de la oportunidad de exponentes como "errores pronominales", "doble negación" o "forma incorrecta de palabra", así como de la capacidad del programa de identificar su recurrencia. Ahora bien, resultan mucho menos convincentes los criterios retenidos, por ejemplo, para medir el estilo, a saber, "repetición de palabras", "frases demasiado cortas", "frases demasiado largas", "demasiadas frases comenzadas con conjunciones" y "uso de la voz pasiva" (Deane, 2013). Parece, así, fundamentalmente discutible que el resultado acumulado de una cadena semejante de valores cuantitativos equivalga de forma alguna a la valoración análoga que del mismo constructo pudiera realizar una persona.

De hecho, Quinlan *et al.* (2010) -que pertenecen a la plantilla de investigadores del ETS- identifican explícitamente los criterios de voz, ideas y contenidos, organización, elección léxica y fluidez de lectura como aquellos que suponen una mayor dificultad para el "e-rater" en particular y las herramientas automatizadas de evaluación en general. Es más, su análisis de dos estudios experimentales diferentes revelan otras debilidades. En primer lugar, toda una serie de categorizaciones inapropiadas dentro de un corpus de textos corregidos, que los autores relacionan con etiquetas inadecuadas introducidas en el momento de calibrar el instrumento. Un aspecto de este tipo, en realidad, apunta a una cuestión más amplia, a saber, la posibilidad muy real de una clasificación fallida de errores como resultado de la ausencia de un proceso interpretativo de los mismos. Algo así resulta particularmente pertinente en contextos de instrucción de una L2, en los que un determinado segmento textual incorrecto puede presentar diferentes niveles de desviación gramátical o léxica con respecto a la norma, de tal forma que resulta necesario un esfuerzo consciente de movilización de la información contextual disponible por parte del lector para reconstruir el significado que se pretendía transmitir. Una operación semejante

constituye prácticamente un acto reflejo para cualquier profesional de la enseñanza o la calificación pero, al depender de toda una serie de inferencias y reinterpretaciones del significado, resulta imposible para una herramienta de carácter automático.

En cualquier caso, volviendo a las conclusiones del estudio, la revisión de los cálculos estadísticos realizados reveló asimismo de forma inesperada toda una serie de fallos relacionados con rasgos formales aislados en principio tan anodinos como la ausencia de puntuación. Aún es más, el análisis cuestionaba de forma general la relevancia de constructo de algunos de los índices calibrados. Así, por ejemplo, el indicador de "voz pasiva" presentaba una correlación positiva con la puntuación global de estilo, mientras que, en opinión de los autores, debía resultar negativa "puesto que los libros de textos frecuentemente recomiendan que se utilice la voz activa y el subrasgo de voz pasiva se desarrolló para medir esta trasgresión de estilo" (p. 20) -todo ello a pesar de la recurrencia de dicha estructura en el estilo textual académico. Si bien esta diferencia de perspectivas puede constituir un debate lingüístico pertinente y enriquecedor desde el punto de vista didáctico, en este caso sirve para custionar lo apropiado de la noción consistente en aplicar una serie de microcriterios de medición a a toda una serie de tareas, independientemente de su género textual. Los autores, por consiguiente, reconocían que el constructo se cubría tan sólo de forma parcial.

Los investigadores del ETS, de hecho, han destacado por encabezar los esfuerzos de investigación centrados en cuestionar la validez del "e-rater" como muestra de transparencia y legitimidad, en abierto contraste con la mayor parte del resto de sistemas de evaluación automatizada, cuyo funcionamiento suelen mantener en secreto sus respectivos vendedores invocando derechos de propiedad. Powers *et al.* (2001), por ejemplo, invitaron a toda una sere de especialistas académicos de disciplinas relacionadas, desafiándolos a que engañaran deliberadamente al sistema escribiendo textos que obtuvieran una calificación mucho mayor o menor de la que objetivamente se merecían. El experimento demostró que resultaba mucho más fácil burlar al dispositivo para conseguir una mejor calificación que de forma inversa, lo que los autores interpretaron como una indicación tranquilizadora de la improbabilidad de una penalización injustificada a los estudiantes. La mayor discrepancia, así, se apreció en un texto compuesto de una serie de parráfos repetidos hasta 32 ocasiones, al cual el "e-rater" otorgó la calificación más alta (6), mientras que el examinador experto le asignó la más baja (1). Así las cosas, los autores admitieron la posibilidad de que algunos intentos malintencionados explícitos consiguieran confundir al programa. No obstante, el experimento reveló que tal logro requería un nivel de conocimiento teórico y complejidad deliberada tales para no poder plantearse como una estrategia rentable al alcance de un candidato común. Su conclusión, en cualquier caso, era que el "e-rater" no estaba en condiciones de funcionar de forma autónoma.

Desde una perspectiva distinta, con el objetivo de evaluar la validez de los exámenes TOEFL realizados a través de Internet y puntuados con el uso de calificadores automatizados, Enright y Quinlan (2010) se centraron en cuatro inferencias comúnmente empleadas en un argumento de validez: la de evaluación, la de generalización, la de extrapolación y la de utilización. Admitieron que el modelo, a pesar de lograr altos índices de predicción de resultados (lo cual satisface el aspecto de generalización), no atendía satisfactoriamente a aspectos centrales del constructo como la eficiencia, la cohesión o la adecuación de la respuesta. En lo que se refería a su capacidad de extrapolación, tras comparar los ejercicios corregidos con muestras

de escritura externas de los mismos candidatos producidas en el contexto de instrucción universitaria, quedó de manifiesto la misma baja correlación que la apreciada con los exámenes corregidos por examinadores, lo cual vendría a cuestionar más bien el propio formato de la prueba. Por último, no se pudo proporcionar evidencia alguna para sostener una inferencia de utilización -relacionada con el uso posterior de los resultados en un contexto educativo-, si bien los autores apuntaron la importancia de atender el impacto en el aula que sucitaría la introducción de correctores automáticos, así como de las consecuencias más generales en el ámbito de la enseñanza.

Un último estudio que citaremos aquí es el de Weigle (2013), que propone algunas consideraciones críticas para la aplicación de herramientas automatizadas de evaluación en contextos de L2. La autora establece una diferencia entre la evaluación de la escritura por un lado, la evaluación del contenido a través de la escritura por otro y, por último, la evaluación de la lengua a través de la escritura. Así, las dos primeras aparecerían como emblemáticas de los contextos de enseñanza en y de L1 y la última de los de L2. Si bien reconoce las objeciones acerca de la incapacidad de este tipo de sistemas para responder de forma satisfactoria a consideraciones de estilo, apela a las diferencias en el constructo de escritura evaluable en contextos de L2, en comparación con la instrucción de expresión escrita en L1 o las asignaturas en las que se evalúa la adquisición de un contenido temático. Así, en su opinión, los exámenes de dominio de lengua como el TOEFL, presentan un contexto más apropiado para la introducción de herramientas de puntuación automática, puesto que las consideraciones de carácter retórico poseen una relevancia más atenuada mientras que se presta mayor atención a cuestiones formales relacionadas con la corrección lingüística. Se trata de una observación pertinente pero discutible a la que volveremos más adelante.

4. CRÍTICAS A LOS SISTEMAS DE EVALUACIÓN AUTOMÁTICA

Pasemos ahora a ocuparnos de los argumentos que los críticos con estos sistemas hacen valer. Resulta, así, particularmente cautivador adentrarse en las demostraciones retóricas articuladas desde la bancada de los detractores frontales, colectivo compuesto en su mayor parte por profesores e investigadores de expresión escrita y lengua inglesa. En estas muestras discursivas de férrea convicción se tiende a escenificar una hostilidad abierta frente a los expertos en medición académica, implícitamente presentados como grises programadores totalmente insensibles a la interacción humana que caracteriza al trabajo en el aula. Sus contribuciones exaltan el papel de los "profesores en trincheras" (Ericsson, 2006) y fustigan los "océanos capitalistas de la calificación de textos automática donde pastan el "e-rater" del ETS (...) y el WritePlacer del College Board" (Haswell, 2006: 57). Herrington llegó incluso a colarse en una presentación comercial del Intellimetric de los laboratorios Vantage y transcribió los comentarios de los participantes, que ensalzaban las herramientas de calificación automatizadas por no cansarse ni sufrir de dolores de cabeza, contrariamente a los docentes. "Consignamos estas observaciones", apunta, "porque proporcionan una perspectiva de la mentalidad de este grupo de gente de la evaluación y los exámenes: nosotros, profesores de inglés, somos presentados como poco fiables e inconvenientes. Somos presentados por otros, de forma más generosa, como resistentes reaccionarios" (Herrington & Moran, 2001:486). En este

artículo, los autores alarman abiertamente acerca de las implicaciones de sustituir a profesores y correctores por máquinas, al identificar no sólo una amenaza a su propia posición laboral, sino también una peligrosa alteración en el mero hecho de escribir, que pasaría a percibirse "como un acto formal de demostración y no una interacción retórica entre alguien que escribe y sus lectores" (481). Este tipo de intervenciones registradas en foros especializados no ocultan una oposición aprioirística y se distancian del formato cuantitativo experimental para adoptar dimensiones cercanas al discurso en el ágora, de notable riqueza retórica.

Una de las críticas enunciadas más destacables es la que propone Anson (2006), quien replica en las primeras páginas de su contribución la cadena de procesos cognitivos sucesivos pero no lineales que realizaría una persona leyendo ese mismo artículo. Acto seguido pasa revista al desarrollo histórico de las aplicaciones de inteligencia artificial al lenguaje natural, vinculando sus fracasos a su absoluta incapacidad para realizar inferencias, reconocer temas o papeles y sus consiguientes connotaciones o "aprender" de aquello que se ha descodificado previamente. No en vano, todas éstas constituirían funciones interpretativas básicas, subyacentes a la propia naturaleza de la lectura. Si bien el autor reconoce explícitamente un espacio para la tecnología informática en el desarrollo de la competencia escrita, advierte con firmeza a propósito de los riesgos que conllevaría el hecho de escribir a máquinas, al señalar que "cuando (los estudiantes) son conscientes de la subyugación de las circunstancias humanas a un "lector" fríamente objetivo que ni piensa, ni siente y que es abiertamente indiferente a la interacción personal -incluso en el contexto de un examen de gran repercursión que de por sí carece en gran parte de tal intención-, la comunicación humana queda relegada al silencio" (54).

Por su parte, Williamson (2004) rechaza las condenas tajantes y recuerda que ninguno de los responsables de las diferentes herramientas afirma que el sistema sea capaz de leer, puesto que leer y puntuar de forma holística constituyen niveles diferentes de procesamiento textual. En su opinión, la controversia debería reconceptualizarse con mayor serenidad en torno a la simple cuestión de si la calificación holística, tal y como se practica en los exámenes de certificación, se puede computar técnicamente, es decir, si resulta posible desarrollar un algoritmo de programación que pueda replicar el mismo proceso de forma más rápida y a menor coste. Deane (2013) comparte esta visión pragmática aludiendo al hecho de que muchas de las oposiciones virulentas a los sistemas automatizados de calificación constituyen en realidad críticas al constructo de escritura implícito en la situación de un examen estandarizado de certificación, independientemente del método aplicado para su calificación.

De hecho, Condon (2013) asume abiertamente esta posición, considerando que la controversia en torno a los programas de calificación automática no constituye sino un mero elemento de distracción que oculta el asunto cardinal, a saber, el hecho de que textos no preparados escritos en 25 minutos sobre temas que no resultan familiares -es decir, el ejercicio tradicional de un examen certificativo- constituyen representaciones muy mediocres del constructo que supone la competencia en expresión escrita, más allá de que la puntuación la realice una persona o se delegue en una máquina. Subraya, así, que la puntuación holística proporciona un artefacto estadístico allí donde se requiere información compleja para desarrollar la instrucción de escritura, con lo que programar ordenadores para que repliquen una operación humana originalmente defectuosa constituye una innovación de dudoso

interés. Es más, el autor diferencia lo que supone calificar la calidad de la expresión escrita de la mera identificación de los aspectos superficiales que se le relacionan -lo que en su opinión realizan las herramientas informatizadas-, sugiriendo de este modo que hasta que un lector automático no esté en condiciones de procesar el contraste semántico entre las frases "El asado está listo para comer" y "El tigre está listo para comer", cualquier novedad que los investigadores relacionados puedan presentar como un avance resultará fundamentalmente irrelevante. Condon concluye así que un conjunto acumulado de rasgos lingüísticos aislados, tal y como se operacionaliza en el constructo de los evaluadores automáticos, difícilmente puede medir o explicar la destreza de producción escrita.

Ericsson (2006) enuncia una deconstrucción similar de la supuesta validez de constructo vinculada a los sistemas de evaluación automática. Vitupera así de forma directa su asunción subyacente, a saber, que un texto redactado por aprendientes puede equipararse a "una bolsa de palabras", considerando, por el contrario, que un párrafo es más que una ecuación. Así, mientras defiende que las disquisiciones relativas al significado y sus implicaciones no tienen nada de esotérico, invoca perspectivas socioculturales de instrucción de escritura para recordar que el sentido no puede aislarse del contexto humano sociocultural en el que se concibió. Concluye así que "los estudiantes que escriben por y para máquinas no desarrollarán ninguna conciencia de las dinámicas de la lengua ni estarán en condiciones de alcanzar una comprensión de las diferentes audiencias posibles y su necesidad a adaptarse a las mismas" (37).

Por último, Chen y Cheng (2008) presentan una investigación experimental que resulta fundamentalmente crítica con el rendimiento de los sistemas de calificación automática. Se centraron, así, en el uso de estas herramientas para evaluar el desarrollo lingüístico desde una perspectiva formativa, en el marco del trabajo de aula. Utilizaron para ello MyAccess!, una herramienta en línea de Vantage que incorpora las puntuaciones de Intellimetric en un contexto de ejercicios de práctica de baja repercusión, proporcionando retroalimentación a través de una escala holística y otra analítica. Se monitorizaron, pues, tres clases diferentes de escritura en L2 en Taiwán en las que se introdujo durante un semestre el uso de MyAcess!, si bien se dejó su uso a criterio del docente, y se analizaron con cuestionarios y entrevistas las impresiones del alumnado. Aunque la valoración de éstos apareció explícitamente condicionada por la manera en la que la herramienta había sido integrada en el trabajo de clase, ninguna de las personas que respondieron se manifestó de acuerdo con la adecuación de las calificaciones, mientras que menos del 25% consideró que la retroalimentación recibida le había servido para mejorar sus borradores. Por el contrario, la mayor parte expresó frustración ante la información formularia, genérica y a menudo ambigua recibida tras escribir sus textos.

5. CONCLUSIONES

Aunque las líneas precedentes puedan transmitir la idea de un campo de batalla descarnado en el que hordas de robots fanáticos de la fiabilidad formal y la rentabilidad se enfrentan a batallones de profesores que esgrimen en trance el estandarte de la auténtica validez de constructo, lo cierto es que buena parte de expertos en la materia se ubica en una línea continua ajena a los planteamientos binarios. Wiliamson *et al.* (2012) presentan así un abanico de posibles implementaciones para las herramientas de evaluación

automática extendidas entre dos extremos, definidos por su uso exclusivo por una parte y la intervención única de calificadores humanos por otra. Quedaría así entre medio una amplia area de aplicación en la que estos sistemas podrían contribuir a la labor de los examinadores sin asumir por ello el volante del proceso. De hecho, si bien los partidarios de los calificadores automáticos resultan perentorios en su percepción de que éstos "han venido para quedarse" (Williamson, 2004; Shermis, 2014), nadie esconde las limitaciones a las que dichas aplicaciones se encuentran actualmente sujetas, ni aboga por el momento por su uso exclusivo en la práctica de aula o en contextos formales de exámenes. Deane (2013), por ejemplo, considera que la verdadera promesa de este tipo de tecnología se revela una vez que se despliega con cierta originalidad, no en relación con certificaciones o exámenes de alta repercursión, sino más bien concetrándose en asistir el proceso de ideas de quien escribe y su desarrollo en una expresión de calidad. Incluso entre los detractores de los evaluadores automáticos, como veíamos, se aprecia espacio para el compromiso. Anson (2006), por ejemplo, insiste en la utilidad de diferentes aplicaciones de tecnologías informáticas en el ámbito de la pedagogía de la expresión escrita, desde la consideración de que hay más territorio por explorar con una perspectiva de evaluación formativa en vez de sumativa.

Resulta claro, en cualquier caso, que el debate ilustra hasta qué punto los expertos en medición educacional y los docentes de composición y redacción pueden hacer valer no sólo prioridades diversas, sino incluso opiniones abiertamente contradictorias sobre los futuros desarrollos en el ámbito. Attali y Powers (2009), por ejemplo, describen un experimento con el "e-rater" dirigido a desarrollar un único estándar de calificación válido para cualquier tipo de tareas, de textos y años escolares de todos los ciclos educativos primarios y secundarios, esto es, una única herramienta para evaluar la progresión a nivel estatal o nacional a lo largo de los Estados Unidos. Una noción semejante aparece como exactamente lo opuesto a aquello por lo que aboga Huot (2002) al describir las pautas que deberían fundamentar los sistemas de evaluación, a saber, basados en el marco inmediato, controlados localmente y sensibles al contexto. Desde esta perspectiva, la búsqueda de la objetividad en la evaluación de la expresión escrita debería percibirse no como una prioridad sino más bien como un peligro, ya que vendría marcada por la pretensión de imponer uniformidad en lo que constituye esencialmente creación e interpretación. Y si bien las implicaciones de tal afirmación suponen un abierto desafío para los contextos de certificación dirigidos a públicos amplios, la premisa resulta pertinente en la medida que ilumina las evoluaciones a menudo contradictorias entre el ámbito de la pedagogía de la escritura por un lado y su evaluación por otro.

Así las cosas, apuntando hacia una reflexión que dote de cierto carácter interpretativo este panorama de la cuestión, sí que cabría identificar toda una serie de impresiones que desde la perspectiva docente justifican la percepción de los correctores automatizados como elementos potencialmente amenazantes. En primer lugar, la perturbadora sensación que permea buena parte del discurso científico de los partidarios de estos sistemas y aplicaciones, rayano en la persuasión del vendedor a domicilio. Burnstein *et al.* (2012) hablan así abiertamente de la importancia del "mercado K-12[3]", Weigle (2013) trae a colación los trece millones de estudiantes de inglés que se presentan anualmente en China al examen lingüístico universitario *ad hoc* (p. 87), mientras que Shermis (2014) describe el

[3] El término alude a la suma de las etapas primaria y secundaria en los Estados Unidos.

contexto de las evoluciones educativas en su país como una oportunidad para incrementar el uso de las herramientas de corrección automatizada. Puede resultar banal o sumamente cándido manifestar aprensiones ante lo que constituye objetivamente un producto comercial, teniendo en cuenta que los negocios de carácter educativo constituyen una parte integrante de la actividad económica, así como un sector notable de las realidades académicas en todo el mundo. No obstante, las reservas ante lo que se presenta fundamentalmente como proyectos centrados en la obtención de beneficios deberían aceptarse como legítimas, así como la voluntad de no asociarse con los mismos.

Una segunda cuestión aparece con particular relevancia desde la perspectiva de quien enseña. Es aquello a lo que Haswell (2006) se refiere como "el fundamento del burro de carga" para legitimar los sistemas de evaluación automática, en virtud del cual estas aplicaciones aparecen como el *deus ex machina* que salva a estudiantes, docentes e instituciones del lastre que supone corregir y puntuar montañas de textos y exámenes. Ahora bien, para cualquier profesional que considere que responder a los textos de sus estudiantes constituye un elemento central en el establecimiento de una relación de confianza y el desarrollo de habilidades de producción, la mera noción de delegar la retroalimentación y la calificación a un aparato automatizado sólo puede percibirse como un desposeimiento y una dejación manifiesta de responsabilidades. Eso sin mencionar que, tal y como apuntan Herrington y Moran (2001), su introducción no resolvería los problemas institucionales que suelen hallarse detrás del profesorado sobrecargado -a saber, clases masificadas y ratios elevadas- sino que más bien favorecería su perpetuación, ya que se proporcionaría una herramienta para gestionar tal realidad. En efecto, con la generalización de los correctores automáticos, "las autoridades administrativas considerarían que todo lo que resultaría necesario sería un único docente en tal institución, o quizás en el Estado o en la nación, para enseñar Biología de primer año o Shakespeare a todos los estudiantes" (p. 496).

Es cierto que Herrington, Moran y sus colegas pertenecen al colectivo académico de Didáctica de la escritura y la composición en L1, que en Estados Unidos posee una presencia transversal entre disciplinas y especialidades. Así, sus declaraciones alertando sobre las consecuencias de delegar sus esfuerzos o los de los docentes de áreas temáticas concretas como la Historia o la Geología en aplicaciones informáticas incapaces de procesar el significado resultan difíciles de ignorar, incluso por parte de los más ardientes defensores de dichos sistemas. También es cierto que la situación retórica artificial que denuncian cuando un trabajo o un comentario de texto se dirige a un calificador automatizado puede no resultar igual de pertinente en un contexto de enseñanza de L2. No en vano, contrariamente a lo que puede ocurrir con una disertación de Literatura, Biología o Ciencias Sociales, las tareas de producción escrita en L2 empleadas en contextos certificativos -al menos cuando se adopta una perspectiva comunicativa- proporcionan una audencia concreta como un parámetro definitorio que permite evaluar el uso de la lengua y su adecuación. Quien se enfrenta a ese examen dirige, pues, sus líneas a este destinatario ficticio y no al calificador, ya sea humano o automático. ¿Habría que concluir, pues, como apuntaba Weigle (2013), que la enseñanza y evaluación de una L2 constituye un campo experimental pertinente para las herramientas de evaluación automática?

La respuesta que en estas líneas se apuntará será en todo caso negativa. La escritura, en tanto que desempeño de carácter comunicativo, tiene que ver fundamentalmente con lo que Berthoff

bautizó como la creación de significado (1978). Para cualquiera de las personas implicadas en el desarrollo de habilidades productivas de expresión escrita y que desempeñen esta función con un mínimo de convicción, el trabajo de aula gira en torno a proponer tareas estimulantes e invitar a estudiantes a compartir experiencias, expresar ideas, crear mundos nuevos. Y para ello un paso fundamental lo constituye inculcar la noción de que un texto -incluso en el contexto de la docencia de una L2 en el que el interés práctico último puede ser efectivamente el desarrollo lingüístico- supone una oportunidad para articular pensamientos e intenciones dentro de un contexto educativo, no sólo un pretexto para revisar tiempos verbales, el uso de preposiciones o las convenciones ortográficas. Resulta totalmente cierto que la evaluación como ámbito va mucho más allá del diseño y corrección de exámenes. También lo es el hecho de que las instancias certificativas globalizadas que constituyen el área natural de aplicación de los sistemas de corrección automática no suelen proporcionar ningún tipo de retroalimentación, no se prestan de forma alguna a planteamientos evaluativos innovadores ni facilitan tiempo para elaborar borradores o plantear revisiones. No obstante, precisamente al excluir estos procesos, pertenecientes a la naturaleza fundamental del hecho de escribir en cualquier lengua y en la mayor parte de contextos, aquellas pruebas certificativas y sistemas de evaluación que se apoyen exclusivamente en tal tipo de tareas improvisadas y ceñidas al cronómetro desfiguran el constructo de la expresión escrita. Parecería pues oportuno preocuparse por enmendar dicha carencia y abogar por cambios que observen una mínima coherencia entre la instrucción contemporánea de la expresión escrita y cómo se evalúa en exámenes certificativos de alta repercursión. Mientras tanto, parece como mínimo una deferencia garantizar a sus candidatos que una persona experta -más allá de las tensiones a las que se vea sometida, de los dolores de cabeza que pueda sufrir o del malestar emocional en el que pueda verse envuelta- se ocupará de leer las líneas que han redactado y que hará lo posible por seguir y recomponer el significado que han creado para la ocasión, incluso, quién sabe, aprobando, indignándose, sonriendo o emocionándose con lo que está leyendo. Y no se trataría aquí tanto de un disgresión sentimental ajena al rigor científico como de una obligación deontológica.

BIBLIOGRAFÍA

Anson, C.M. (2006). Can't Touch This: Reflections on the Servitude of Computers as Readers. En Ericsson, P.F., & Haswell, R.H. (eds.), *Machine scoring of human essays*, *Truth and consequences*. Logan: Utah University Press.

Attali, Y. (2007). Construct validity of e-rater® in scoring TOEFL® essays. *ETS Research Report Series*, 1, i-22.

Attali, Y., & Powers, D. (2009). Validity of scores for a developmental writing scale based on automated scoring. *Educational and psychological measurement,* 69(6): 978-993.

Berthoff, A.E. (1978). Tolstoy, Vygotsky, and the making of meaning. *College Composition and Communication*, 29(3): 249-255.

Chen, C.F.E., & Cheng, W. Y. E. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived

learning effectiveness in EFL writing classes. *Language Learning & Technology*, 12(2): 94-112.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing*, 18(1): 100-108.

Crossley, S.A., Kyle, K., Allen, L., Gou, L., & McNamara, D.S. (2014). Linguistic microfeatures to predict L2 writing proficiency: A case study in Automated Writing Evaluation. *The Journal of Writing Assessment*, 7(1): 116-135.

Crossley, S.A., & McNamara, D.S. (2012). Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2): 115-135.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, 18(1): 7-24.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1): 1-35.

Enright, M.K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, 27(3): 317-334.

Ericsson, P.F. (2006). The meaning of meaning: Is a paragraph more than an equation. En Ericsson, P.F., & Haswell, R.H. (eds.), *Machine scoring of human essays*, *Truth and consequences*. Logan: Utah University Press.

Haswell, R. (2006). Automatons and automated scoring: Drudges, black boxes, and dei ex machina. En Ericsson, P.F., & Haswell, R.H (eds.), *Machine scoring of human essays*, *Truth and consequences*. Logan: Utah University Press.

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English*, 63(4): 480-499.

Huot, B. (2002). *(Re) articulating writing assessment for teaching and learning*. Logan: Utah University Press.

Messick, S. (1993). Foundations of validity: Meaning and consequences in psychological assessment. *ETS Research Report Series*, 1993(2): 1-18.

Powers, D.E., Burstein, J.C., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). Stumping E-Rater: Challenging the validity of automated essay scoring. *ETS Research Report Series*, 2001(1): 1-44.

Quinlan, T., Higgins, D., & Wolff, S. (2009). Evaluating the construct-coverage of the e-rater® scoring engine. *ETS Research Report Series*, 2009(1): 1-35.

Shermis, M.D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20: 53-76.

Weigle, S.C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, 18(1): 85-99.

Williamson, M. (2004). Validity of automated scoring: Prologue for a continuing discussion of machine scoring student writing. *Journal of Writing Assessment*, 1(2): 85-104.

Williamson, D.M., Xi, X., & Breyer, F.J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1): 2-13

FECHA DE ENVÍO: 13 DE NOVIEMBRE DE 2016