ISSN: 1139-613X

5

INTRODUCCIÓN A LA EVALUACIÓN DE TEXTOS POR ORDENADOR EN LA ENSEÑANZA DE UNA LENGUA EXTRANJERA. EL PROGRAMA ESSA.

(INTRODUCTION TO COMPUTER ESSAY EVALUATION IN FOREIGN LANGUAGE TEACHING. ESSA SOFTWARE)

José Luis García González
Universidad de Cantabria

RESUMEN

La evaluación automatizada de textos es un campo de investigación emergente que está demostrando el potencial que los ordenadores tienen a la hora de procesar el lenguaje y de ser aplicados a la evaluación de textos. Este artículo ofrece una introducción a los principales conceptos y procedimientos empleados en el diseño de software para evaluar composiciones escritas. ESSA es un programa fruto de las investigaciones en este campo. Describiremos su funcionamiento para comprender mejor cómo analiza y evalúa narraciones cortas en el área de inglés. El grado de acuerdo entre la calificación de ESSA y la de cuatro docentes evaluando los mismos textos ascendió a 0.799.

ABSTRACT

Automated essay scoring is an emerging research field which is showing the potential computers have when processing natural language as well as in marking essays. This article offers an introduction to the main concepts and procedures used when designing software for scoring texts. ESSA software is one of the products in this area of research. We describe its operational design in order to better understand how ESSA analyzes and grades short stories in a foreign language class. The final results between ESSA and four teachers marking the same texts was 0.799.

INTRODUCCIÓN

La evaluación automatizada de textos tiene su origen en 1966, cuando Ellis Page desarrolló por primera vez un programa que llamó *Project Essay Grade*™ (PEG). Desde entonces han sido múltiples los intentos de elaborar programas con el objetivo de que pudieran ser utilizados por los docentes para valorar las composiciones escritas tanto en pruebas de nivel como en los ejercicios de aula tradicionales en el sistema anglosajón (Haswell, 2006).

El campo de investigación sobre análisis y recuperación de la información se ha extendido a diversas facetas. El objetivo es conseguir un nivel de ejecución suficiente para que los ordenadores analicen y procesen la ingente cantidad de información que actualmente se genera en forma de textos. Ese cometido empieza a ser posible por las ventajas que poseen a la hora de realizar cálculos complejos así como por su facilidad para extraer unidades de análisis de los casi infinitos documentos que se pueden llegar a generar.

Estamos hablando de la disciplina científica denominada Inteligencia Artificial y el desarrollo a su amparo de lo que se conoce como Procesamiento del Lenguaje Natural (NLP, siglas en inglés) y la Lingüística Computacional.

Las TIC cada día están más presentes en nuestro entorno. En educación también empieza a consolidarse su papel y a reconocerse su valor. El uso de las TIC se centra en el proceso de enseñanza y aprendizaje, quedando todavía al margen la evaluación educativa. Se trata de complementar la evaluación tradicional con pruebas estructuradas que se alejan de la realidad lingüística que pretenden medir.

Es necesario diseñar actividades basadas en el currículo y un sistema de evaluación implementado consecuentemente (Myers, 2003). Se podrían medir las cuestiones gramaticales con tests; pero la gramática no es un objetivo en sí mismo, sino su aprendizaje contextualizado. Las pruebas objetivas verifican un conocimiento aislado, que, en muchos casos, no es suficiente para demostrar que el estudiante es capaz de utilizarlo cuando tiene que enfrentarse a una situación comunicativa.

1. SOFTWARE PARA EVALUAR TEXTOS

El software para evaluar textos ha tenido considerable repercusión educativa en los países anglosajones. En ellos se inició la investigación y es

donde más popularidad han alcanzado. Los programas que sobresalen actualmente en este campo son los siguientes: IEA™, PEG™, MY ACCESS!®, ETS™ y BETSY (Dikli, 2006; Rudner y Gagne, 2001).

Proyect Essay Grade™ (PEG™)

PEG™ fue el primer programa desarrollado para valorar el estilo y contenido de textos escritos para una prueba de admisión GMAT, *Graduate Management Admission Council,* por alumnos graduados en Estados Unidos.

Page concluye que los ordenadores tienen un alto potencial para igualar o superar la fiabilidad de los docentes en las evaluaciones, con las siguientes ventajas:

1. Los resultados experimentales superaron a las puntuaciones dadas por dos jueces.
2. Los textos son corregidos con mayor celeridad, al haber menos personas implicadas.
3. La evaluación automatizada supone un ahorro del 97% en los costes.
4. Para las valoraciones individuales, las puntuaciones son más descriptivas que las calificaciones proporcionadas por dos jueces.
5. Procedimientos para verificar la validez pueden ser introducidos en el programa y corregir así posibles desviaciones humanas o de las máquinas (Page, 2003).

El programa utiliza entre 30 y 40 variables a la hora de valorar los textos (Page, 2003). PEG™ se basa en el modelo de regresión múltiple y utiliza como variables independientes aspectos cuantificables en un texto, tales como: longitud del texto, longitud de las palabras y puntuación (Rudner y Gagne, 2001). Valenti et al. (2003) añaden más procesos que PEG™ realiza al analizar un texto:

> Cuenta las preposiciones, pronombres relativos y otras partes de la oración como un indicador de la complejidad de la estructura oracional. La longitud de las palabras indica la dicción (pues las palabras menos comunes son más largas). [...] PEG también requiere entrenamiento, valorando un cierto número de textos con las variables independientes previamente marcadas manualmente, para poder obtener los coeficientes de regresión, que a su vez permiten evaluar nuevos textos (Valenti et al., 2003, 320-321. Los paréntesis proceden del original).

PEG™ integra una serie de herramientas como un corrector ortográfico, un diccionario y un analizador de partes de la oración. La correlación entre PEG™ y los evaluadores ha sido r = 0,87 (Valenti et al., 2003). El programa evalúa y proporciona *feedback* sobre los siguientes aspectos del texto: ideas, organización, estilo, corrección gramatical y creatividad.

Intelligent Essay Assesor™ **(**IEA™)

IEA™ fue creado por Thomas Landauer y Peter Foltz. Se patentó por primera vez en 1989 (Rudner y Gagne, 2001). El programa crea índices de los documentos y luego emplea la técnica denominada análisis semántico latente.

El grado de acuerdo entre IEA™ y las puntuaciones dadas por unos jueces, expresado en porcentaje, oscila entre el 85 y el 91% (Valenti et al., 2003). La correlación entre IEA™ y los docentes fue del 0.701%.

La información que IEA™ proporciona, tras efectuar el análisis, incluye cuatro aspectos: audiencia y objetivo, estructura, argumentación y uso de la lengua.

Electronic Essay Rater® (E-Rater®)

E-Rater® fue creado por el *Educational Service Testing* ('servicio norteamericano de tests educativos') y utilizado para su cometido desde 1999. Rudner y Gagne (2001) describen el funcionamiento básico de E-Rater®, indicando varias características comunes con PEG™, aunque incluye otros procedimientos estadísticos más complejos:

> Es una «tecnología híbrida del futuro», pues se basa en el análisis de la variedad sintáctica, estructura del discurso (como PEG™) y en el análisis de contenido (como IEA™). Para medir la variedad sintáctica E-rater cuenta el número de complementos, subordinadas, infinitivos, cláusulas de relativo y verbos modales («would, could») para calcular ratios de estas características por oración y por texto.
> (En http://pareonline.net/getvn.asp?v=7&n=26. Parénteis y comillas proceden del original)

Mide la variedad sintáctica con una técnica que se denomina procesamiento del lenguaje natural (PLN), para extraer las unidades lingüísticas del texto que serán comparadas con un grupo de textos valorados previamente por docentes. Los factores que considera en su análisis son los siguientes:

— Análisis del contenido
— Nivel de expresión
— Proporción de errores gramaticales
— Proporción de errores de uso
— Proporción de errores expresivos
— Proporción de comentarios estilísticos
— Organización y desarrollo de la puntuación
— Longitud del ensayo

E-Rater® también necesita ser entrenado con, al menos, 300 textos modelo. El programa ha llegado a obtener una correlación con los docentes del 97% (Valenti et al., 2003).

IntelliMetric™

Este programa fue comercializado por el grupo corporativo Vantage Learning en 1998. Es uno de los pocos programas que proporciona *feedback* en 20 lenguas, entre ellas el español.

Al igual que los anteriores programas, la tecnología que emplea está protegida por patentes y el secreto comercial. Se dispone de muy poca información sobre los procesos y técnicas que el programa emplea. Se basa en la inteligencia artificial, el procesamiento del lenguaje natural y técnicas estadísticas (Elliot, 2003), además de lo que indica el grupo Vantage Learning en la publicidad de su producto en la web: tecnologías *CogniSearch*™ y *Quantum Reasoning*™.

IntelliMetric™ necesita ser entrenado con unos 400 relatos. No obstante, sus creadores sostienen que en algún caso ha sido suficiente con una muestra de 50 textos para obtener calificaciones aceptables (Dikli, 2006). El programa proporciona los resultados según las siguientes categorías: punto de vista y coherencia, organización, elaboración y desarrollo, estructura oracional y corrección gramatical (Dikli, 2006).

My Access!™

Se basa en Intellimetric™. Proporciona puntuaciones instantáneas y *feedback* de diagnóstico, para que los usuarios mejoren su composición mientras la elaboran, antes de ser evaluada. Al igual que Intellimetric™, es multilingüe. Otra característica es que proporciona *feedback* multinivel, es decir, según tres niveles: inicial, medio o avanzado (Dikli, 2006), sugiriendo más de doscientos comentarios sobre el análisis del texto.

El programa requiere ser entrenado con unos 300 relatos. El grupo Vantage Learning aduce que el programa puede proporcionar *feedback* sobre textos redactados según diferentes géneros literarios (Philips, 2007).

En cuanto a los resultados, el grupo Vantage Learning afirma que Intellimetric™ alcanza una correlación Pearson promedio de r = 0,93 con la puntuación de los jueces (Vantage Learning, 2006). La revisión de más de 150 estudios que el grupo ha llevado a cabo sobre el grado de ejecución del programa, ofrece las siguientes conclusiones:

1. El grado de acuerdo con la puntuación de expertos, a menudo excede la propia de los expertos entre sí.
2. Califica con precisión respuestas abiertas en una variedad de niveles, áreas y contextos.
3. Demuestra una fuerte relación con otras pruebas que miden el mismo constructo.
4. Evidencia unos resultados estables en todas las muestras. (Vantage Learning, 2007)

Bayesian Essay Test Scoring System (BETSY)

Lawrence Rudner desarrolló BETSY. Se basa en modelos bayesianos propuestos por Naïve Bayes, como el modelo multivariado de Bernoulli y el modelo multinominal, que ya se han mostrado eficaces en tareas como la clasificación de currículos vitae, determinación de noticias sindicadas interesantes para el usuario y la discriminación de la publicidad no deseada en el email, por ejemplo.

Las variables que la evaluación bayesiana emplea son múltiples, e incluyen las ya vistas en programas como PEG™, el análisis semántico latente y E-Rater®. El objetivo es clasificar los textos en una escala categórica o nominal. Como ya manifiesta su autor, en teoría este modelo de evaluación por ordenador disfruta de las ventajas de las mejores características de PEG™, LSA, y E-rater® además de varios atributos específicos cruciales.

BETSY necesita unos 1000 textos por categoría para la que vaya a ser entrenado. El grado de acuerdo con las puntuaciones de los evaluadores ha superado el 80% (Dikli, 2003).

2. LOS CONCEPTOS DE VALIDEZ Y FIABILIDAD EN LA EVALUACIÓN POR ORDENADOR

Fiabilidad y validez son términos que pueden llegar a usarse indistintamente cuando se habla de la evaluación automatizada por ordenador (Cizek y Page, 2003). Dado que los programas EAT (Evaluación Automatizada de Textos) se calibran y validan en función del juicio emitido por los docentes, se parte del hecho de que la evaluación de los docentes es válida y refleja las características de aquello que se quiere medir. Hasta tal punto están relacionados ambos conceptos que la "fiabilidad y validez de las puntuaciones EAT dependen de la correlación que existe entre los jueces participantes en la calibración, y los coeficientes de validez variarán dependiendo de la correlación entre jueces durante la validación" (Keith, 2003, 150).

En el marco de la tecnología EAT, los procedimientos habituales para determinar la fiabilidad de una prueba no se muestran útiles. "La fiabilidad se ha medido tradicionalmente en términos de correlación bien dividiendo la muestra en dos partes bien mediante el procedimiento test-retest. En cualquier caso, ambos métodos producirían los mismos resultados cuando la evaluación es automatizada (Cizek y Page, 2003).

Para medir la fiabilidad y la validez se emplean otros dos procedimientos. Uno consiste en calcular la media de las correlaciones entre el ordenador y cada evaluador. El otro calcula la calificación promedio de los docentes, y su correlación con el ordenador. El segundo es efectivo a la hora de reducir el impacto que las valoraciones extremas tienen en la nota final (Keith, 2003).

Previamente al cálculo de la media de las puntuaciones, hay que revisar si se han producido puntuaciones extremas que puedan afectar a los resultados. Es decir, cuando varios docentes califican un texto, esa valoración pueden hacerla con cierto grado de acuerdo. Cuanto más acuerdo se produzca en sus juicios, más fiabilidad y validez tendrá el modelo que se base en esas puntuaciones. Cizek y Page (2003) presentan dos métodos para determinar el grado de acuerdo: exacto o adyacente. Se adoptará el acuerdo exacto cuando se requiera que las puntuaciones, habitualmente de dos docentes, sean idénticas. Se optará por el acuerdo adyacente, cuando se tolere un margen de diferencia. En el caso que no haya acuerdo, se dan dos opciones: una, eliminar la nota extrema e introducir la que proporcione un evaluador adicional para estos casos. La otra es añadir una valoración extra y calcular la media de todas.

Los principales errores que un programa de estas características puede cometer son dos: falsos positivos y omisiones (Leacock y Chodorow, 2003).

Los falsos positivos son unidades que se identifican erróneamente, y que no deberían ser consideradas como tales. Una omisión es una unidad que debería reconocerse, pero que el programa no identifica.

Más específicamente sobre la validez, tradicionalmente, se consideran tres tipos a la hora de verificar los resultados de unas pruebas: validez de contenido, validez de constructo y validez de criterio (Field, 2005). La validez de contenido no es particularmente relevante cuando se trata de la evaluación automatizada, ya que "se aplica a los textos en sí mismos, más que al programa de evaluación" (Keith, 2003, 148).

La validez de constructo implica utilizar técnicas estadísticas avanzadas como el análisis de regresión múltiple, el análisis de varianza y el análisis factorial. Todos sabemos lo que es evaluar y a lo largo de nuestro período estudiantil hemos pasado por tantas pruebas, que nos consideramos capaces de reproducir esas situaciones. No obstante, determinar qué y cómo se va a evaluar no es una tarea tan sencilla.

Con respecto a la validez de criterio, los programas EAT mantienen una elevada correlación con las puntuaciones dadas por evaluadores humanos (Dikli, 2006), lo que evidencia su potencial a la hora de utilizarlos para calificar.

3. EL PROGRAMA ESSA

La propuesta que presentamos en este artículo se denomina ESSA, acrónimo en inglés de *English Short Story Analyzer*. Está accesible gratuitamente en Internet en la siguiente URL: http://essa.coconia.net.

ESSA fue creado en un principio como un instrumento que ayudara en la identificación de unidades lingüísticas del currículo que el alumnado no nativo empleaba en las composiciones escritas cuando narraba las historias que se trabajaban previamente en el aula, proporcionando además *feedback* instantáneo sobre las mismas. Posteriormente, el programa se modificó para que pudiera, además, evaluar el uso que hacían de aquellos recursos en las narraciones.

El uso del ordenador para evaluar las composiciones, inmediatamente se reflejó en la actitud del estudiante. Su utilización en el aula produjo resultados muy positivos: las composiciones escritas eran para ellos un reto que una máquina iba a valorar. Escribían las historietas, justo después de pulsar en *valorar*, obtenían indicaciones sobre algunos aspectos gramaticales y la calificación. Si la nota era baja, automáticamen-

te repasaban la composición y realizaban correcciones para ver si eso mejoraba la puntuación.

Ello llevó a continuar mejorando el algoritmo que utiliza el programa ESSA. Seguidamente lo describimos de manera general, para ofrecer una idea básica sobre su diseño y funcionamiento.

Primeramente, el programa prepara el texto eliminando o sustituyendo todos los caracteres que no se ajusten al alfabeto anglosajón. En una segunda fase, el programa procesa la información, identificando las unidades gramaticales existentes. ESSA sigue el orden presentado en el diagrama 1 cuando realiza el análisis, para evitar contabilizar unidades más de una vez o unidades compuestas como si fueran la suma de varias simples. A medida que las localiza las elimina del texto, de tal manera que no interfieran con la localización de otras unidades posteriores.

Para realizar el análisis lleva a cabo 7 procedimientos, mediante los cuales localiza:

1. Conjunciones coordinadas
2. Estructuras gramaticales y verbales
3. Verbos habituales
4. Conjunciones subordinadas
5. Otras categorías gramaticales
6. Repeticiones
7. Palabras únicas

Como se aprecia en la diagrama 1, ESSA empieza contabilizando las conjunciones coordinadas y, después, suprimiéndolas del texto. El texto sin las conjunciones coordinadas es el que se utiliza para reconocer el primer grupo de estructuras gramaticales, *estructuras_1*. Se entiende por *estructuras verbales o gramaticales* aquellas construcciones lingüísticas en las que se implican dos o más palabras que tienen entre sí alguna relación gramatical. La dificultad para el discente es obvia desde el momento en el que hay que tener en cuenta la vinculación que pueda existir entre dos o más unidades a la hora de generar una estructura. Ese aspecto ha llevado a considerar un factor específico para las estructuras, que además recoja los contenidos gramaticales. Estas estructuras se localizan en un texto aparte, porque implica eliminar oraciones completas, y el texto resultante carecería de muchas palabras que todavía no se han reconocido.

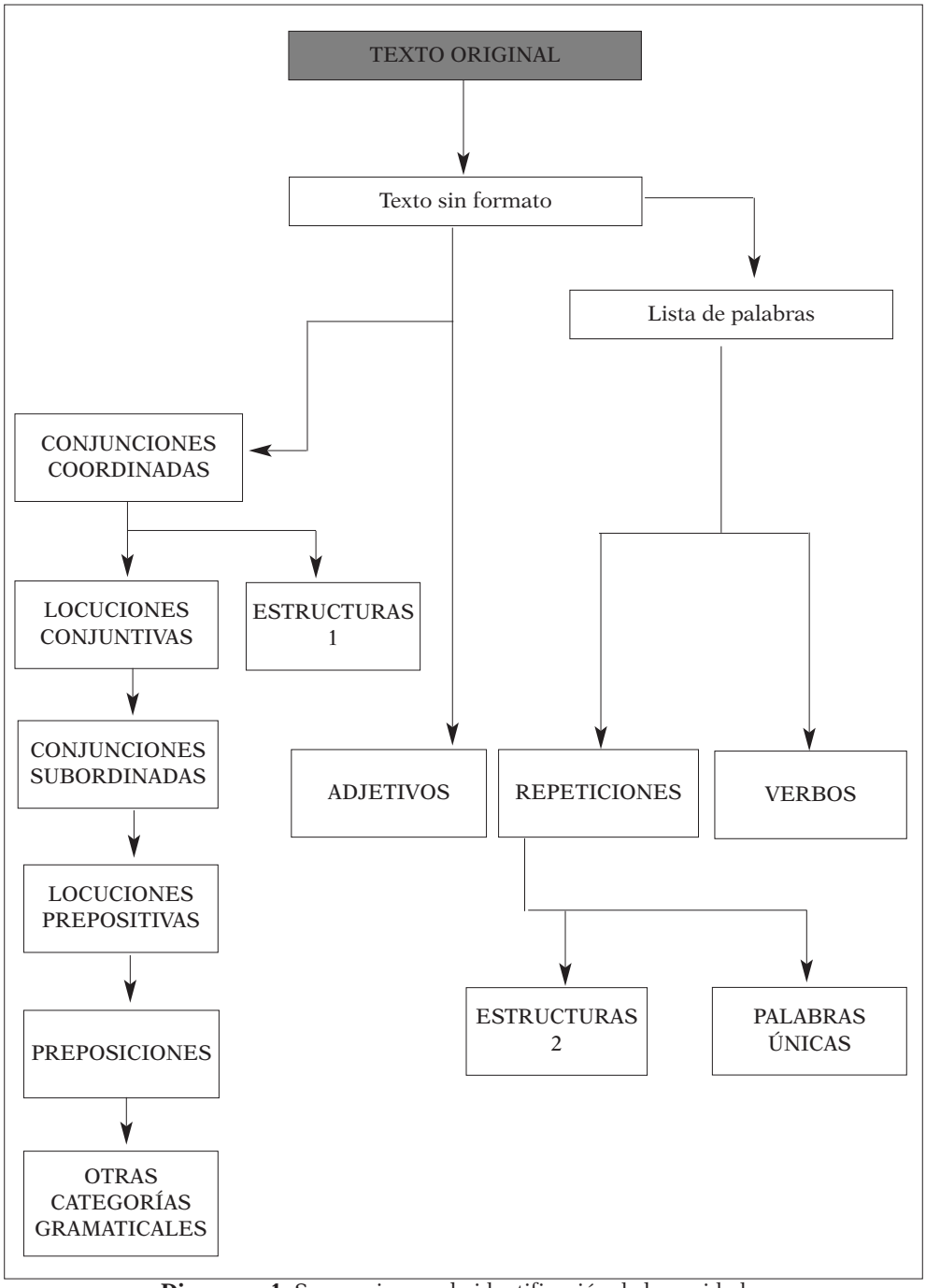

Diagrama 1. Secuencia para la identificación de las unidades

Las estructuras a las que hace referencia este grupo son las que se enumeran seguidamente, respetando el orden indicado:

1. Concordancia pronombre-verbo en tercera persona del singular.
2. Uso de pronombres/adverbios interrogativos en preguntas.
3. Oraciones interrogativas en presente.
4. Oraciones interrogativas en futuro.
5. Oraciones interrogativas en pasado.
6. Formas contractas en presente, incluido el caso posesivo.

7 a 10 Variables vacías para ampliación de estructuras.

El procedimiento de PHP que lleva a cabo para identificar las *estructuras_1* utiliza un lenguaje de búsqueda de cadenas de caracteres denominado *regex* (del inglés, *regular expressions*). Regex es un módulo que incluyen los lenguajes de programación, muy útil cuando se trabaja con texto escrito.

Una vez localizadas las *estructuras_1*, elimina del texto sin formato los adjetivos que en inglés terminan en *-ed/-ing*, para evitar que sean reconocidos como si fueran participios o gerundios. La función *$search_nonverbs* contiene una cadena de búsqueda *regex* con el tipo de adjetivos que se van a suprimir. Cuando son localizados, la función *$replace_nonverbs* los reemplaza por un carácter identificativo.

Una vez eliminados del texto las formas que pueden generar confusión con los verbos, el programa continúa con la localización del grupo de estructuras gramaticales denominado *estructuras_2*, de la misma manera con la que se procedió en el grupo de *estructuras_1*. Las estructuras de este segundo grupo son las siguientes:

11. Las formas de los verbos *have* y *can* en presente.
12. Las formas de los verbos *have* y *can* en pasado.
13. La forma negativa del verbo *have/has*
14. Formas verbales en gerundio o participio con valor de adjetivo
15. La forma *have/has to*
16. Oraciones negativas en pasado
17. La forma *to be going to* en presente
18. La forma *to be going to* en pasado
19. Oraciones negativas en presente continuo
20. Oraciones negativas en pasado continuo
21. Verbos seguidos de infinitivo

22. Verbos seguidos de gerundio
23. Verbos seguidos de infinitivo o de gerundio
24. Primera persona plural del imperativo
25. Oraciones negativas en presente
26. Oraciones negativas en futuro
27. Futuro simple
28. Comparación de adjetivos
29. El grado superlativo
30. Verbos seguidos de un objeto e infinitivo
31. La forma *had to*
32. El presente continuo
33. El pasado continuo

El programa identifica las cadenas de búsqueda una a una con todas las unidades y sus variantes morfológicas, expresadas con patrones de búsqueda *regex* que simplifican todo el proceso, al no tener que enumerar la infinidad de combinaciones que surgirían. Para evitar conflicto con las *estructuras_1*, previamente marcadas en el texto con números correlativos hasta el 10, los indicadores de estas estructuras comienzan a partir del número 11.

El reconocimiento de las estructuras también se lleva a cabo secuencialmente, para evitar solapamientos entre unas y otras. Por ejemplo, busca la expresión del futuro inmediato *be going to* antes que la expresión *verbo + infinitivo*, ya que si no podría interpretar que existen dos estructuras: *be going to* y *going to*. Una vez que identifica las unidades según el orden establecido, las elimina para que no influyan en las restantes. El orden de búsqueda responde al número de unidades que integra cada estructura, comenzando por las más complejas.

Cada estructura lleva asociada una identificación específica, lo que permitirá saber cuántas y cuáles son las que aparecen en el texto. Asimismo, llevan asignado un peso parcial que viene determinado por su complejidad gramatical. Las estructuras más sencillas son las primeras que se trabajan en el aula; mientras que las más complejas, se introducen con posterioridad. Según ese criterio, se valora más el pasado que el presente o el futuro, y las oraciones negativas e interrogativas más que las afirmativas, por ejemplo.

El diagrama 2 muestra el procedimiento seguido para el cómputo de los diferentes factores considerados al calcular la calificación final del texto.

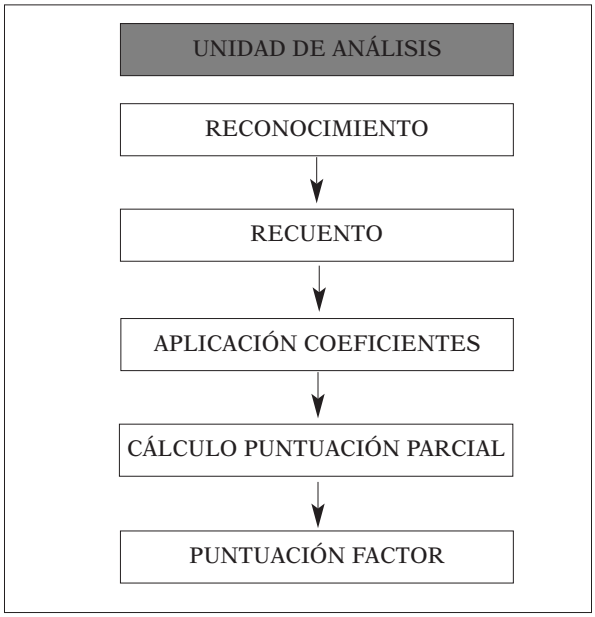

Diagrama 2. Procedimiento para el cómputo de los factores.

El siguiente paso consiste en localizar los verbos. ESSA reconoce los verbos regulares y los irregulares, independientemente del tiempo, modo y persona en el que se usen. Incorpora unos 950 verbos, entre los que se incluyen los verbos regulares habituales hasta el primer ciclo de la ESO, así como la casi totalidad de los irregulares. Además se ha implementado una función en PHP para identificar todas las variantes que pueden presentar en pasado, participio, gerundio o en tercera persona del presente.

Los verbos se han agrupado en tres categorías. Cada categoría lleva asociada un coeficiente. Los verbos en pasado y, especialmente, los irregulares son los que tienen los coeficientes más altos, por la dificultad que representa conocer sus formas.

El texto ideal sería aquél en el que se cuenta una historia utilizando el tiempo pasado. Al redactar en pasado, fácilmente pueden equivocar el tiempo de los verbos. La dificultad para expresarse en tiempo pasado es evidente y, muchas veces, eso se debe a que conocen teóricamente las reglas de formación del pasado de los verbos en inglés, pero falla la aplicación práctica de las mismas. En este caso, el programa verifica esa circunstancia e informa en los resultados sobre una posible confusión de tiempos verbales. ESSA presenta un mensaje de advertencia en la pantalla de los resultados.

El reconocimiento de las conjunciones subordinadas lo realiza a partir del texto del que han sido eliminadas la conjunciones coordinadas, como se señala en el diagrama 2. Primeramente, identifica las locuciones, y después las conjunciones individuales. Esa misma secuencia sigue a la hora de identificar las preposiciones.

El factor O*tras categorías gramaticales* incluye palabras pertenecientes a las siguientes partes de la oración: pronombres, determinantes, adverbios y adjetivos. La dificultad para reconocer adecuadamente su función, ha llevado a aglutinarlas en un solo factor. Por ejemplo, Las siguientes palabras, frecuentes por otra parte, pueden tener varias funciones sintácticas.

— *That*: pronombre, determinante y conjunción
— *What:* pronombre y determinante
— *When*: preposición, adverbio y conjunción
— *Where*: adverbio y conjunción
— *Whose*: pronombre y determinante

Se trata de localizar el término en el texto para comprobar que está presente en el vocabulario del alumnado. La función o el significado que esas palabras poseen no es relevante desde el punto de vista de este análisis.

El alumnado suele cometer repeticiones, aparte de tender a utilizar palabras comodín por su limitado vocabulario. Las palabras repetidas las muestra en la pantalla de los resultados con el número de veces que se han repetido. Se han tenido en cuenta las palabras cuya frecuencia en cualquier texto es alta, para no informar de repeticiones necesarias.

La variable *palabras únicas* muestra la riqueza léxica del texto, en función del número de palabras únicas empleado. Se expresa como la relación entre el número de palabras sin repetir y el total de palabras de la composición. Dos funciones de PHP consiguen, primero agrupar todas las palabras del texto inicial sin formato (*array_count_values)* y, después, contar los agrupamientos para evitar las repeticiones (*count*). Un factor global denominado *vocabulario* valora conjuntamente las variables conjunciones, preposiciones, otras categorías gramaticales y la longitud de la composición.

ESSA ofrece en la pantalla de los resultados del análisis una relación de las unidades reconocidas, para los factores estructuras verbales y gramaticales y verbos habituales. En el primer caso, aparecen todas las estructuras identificadas por su denominación. En el segundo, las categorías a las que pertenecen los verbos localizados. Se ha optado por ofrecer la denomi-

nación de la unidad, en vez de la unidad en sí, por las dos razones siguientes: porque ESSA trabaja con unidades únicas y para evitar la confusión que puede generar una lista de unidades fuera de contexto. Únicamente en el factor *palabras repetidas* indica las repeticiones con el número de veces correspondiente, en el caso de que aparezcan.

Finalizado el análisis del texto, ESSA calcula la valoración y la muestra en pantalla. El programa muestra a los estudiantes, además, las estructuras identificadas, así como el tipo de verbos y su tiempo verbal. También ofrece una valoración cualitativa para cada apartado, de forma que el estudiante puede valorar el uso que hace de los recursos lingüísticos en su composición.

Los mensajes que ESSA presenta en inglés en la pantalla de los resultados, notifican las estructuras gramaticales y formas verbales identificadas en el texto.

El tipo de verbos y el tiempo verbal se indica con cuatro tipo de comentarios, que informan sobre si la redacción está en presente, pasado y si se han empleado verbos regulares y/o irregulares.

— *Regular verbs in the present tense, infinitive or gerund forms*
— *Regular verbs in the simple past or past participle*
— *Irregular verbs in the present tense, infinitive or gerund forms*
— *Irregular verbs in the simple past or past participle*

Finalmente, se muestran tres tipos de mensajes en relación al uso de los tiempos verbales. El siguiente para indicar que la composición se ha realizado en presente y es coherente temporalmente hablando:

— *Verb tense throughout the story seems to be consistent (the present tense)*

Cuando se mezclan distintos tiempos verbales, ESSA presenta uno de los dos siguientes mensajes:

— *Although verb tense throughout the story seems to be consistent (the past tense), you should check if shifts in tense can cause confusion.*
— *Verb tense throughout the story could be inconsistent. Check if shifts in tense can cause confusion*

En esas frases, primero se informa de que el tiempo verbal parece consistente a lo largo de la composición, pero se recomienda repasar la re-

dacción para asegurarse. El segundo indica que la mezcla de tiempos verbales es frecuente y, por lo tanto, pueden generar confusión.

El apartado *vocabulario* engloba los factores preposiciones, conjunciones, otras categorías, longitud de la composición y palabras únicas. La valoración también la indica siguiendo el código comentado. En este caso, los símbolos se han hecho corresponder con una escala que se ajusta según la puntuación global conjunta.

ESSA proporciona la valoración con un número en porcentaje, según una ecuación de regresión, que sustituye a la calificación tradicional, si bien sólo hay que dividir por 10 para obtener la calificación equivalente en nuestro sistema. El hecho de poner la nota en porcentaje se debe al sistema anglosajón, reforzando así el aspecto cultural que se suele trabajar en la clase de idioma.

La calificación final, se guarda en la variable *$nota*, por lo que ya sólo es necesario establecer el rango para cada color: *good,* en verde, si oscila entre 65 y 100 puntos; *fair,* en amarillo, si está entre 41 y 64 y, *poor,* en rojo, si está por debajo de 40 puntos. Cuando la calificación es inferior al 40%, solo aparece la valoración cualitativa.

4. DIFERENCIAS ENTRE LOS PROGRAMAS EAT Y ESSA

Después de la breve panorámica sobre los programas EAT ofrecida y la descripción de ESSA, presentamos algunas diferencias en cuanto a sus características:

1. La principal diferencia es que ESSA ha sido concebido para valorar el uso de los recursos lingüísticos trabajados en el aula que el estudiante utiliza en su composición. Los programas de evaluación automatizada existentes evalúan el nivel de expresión escrita en general y empiezan también a considerar el contenido. Con respecto a la valoración del contenido de los textos, parece algo precipitado confiar en las máquinas la evaluación de tal aspecto, ya que está en una fase inicial y requiere más investigación (Ericsson, 2006; Bird *et al.,* 2009).
2. ESSA incluye las unidades de análisis y los criterios de evaluación referidos al género narrativo que analiza: el cuento. Los programas EAT han de ser entrenados con centenares de textos, previamente calificados, para poder establecer los parámetros estadísticos de la evaluación, ya que su aplicación incluye más géneros literarios y también el rango de la edad del alumnado implicado es mayor.

3. Los relatos analizados por ESSA han de ser creados por alumnado no nativo y, por tanto, el conocimiento de los recursos lingüísticos se limita a los que se concretan en el currículo. La tecnología EAT se enfoca principalmente a la valoración de textos producidos por escritores nativos.
4. Los programas EAT orientan el *feedback* a la mejora de la expresión escrita. En la calificación final, la longitud de las composiciones y la corrección ortográfica tienen un peso significativo. Dado que la extensión de los textos se ha limitado en ESSA y la revisión ortográfica se propone hacerla con anterioridad a la evaluación, estos factores no repercuten en la calificación. En ese sentido, el programa ESSA se centra en identificar los logros del estudiante, en vez de los errores.
5. ESSA proporciona los resultados en una escala entre cero y cien. Los programas EAT operan mayoritariamente sobre una escala que oscila entre cero y seis puntos. Si bien, la escala que se utilice es indiferente, sí lo es el hecho de que la evaluación de ESSA es analítica y cuantitativa; mientras que en los programas EAT es holística y cualitativa.
6. Los programas EAT suelen enfocarse, principalmente, a la evaluación de textos argumentativos. La argumentación requiere unas destrezas superiores a la hora de enfrentarse a un escrito, que implican una capacidad elevada de análisis, de síntesis, de evaluación, de organización de las ideas, etc. Cuando se está en las primeras fases del aprendizaje de una lengua extranjera, esas habilidades quedan fuera de los objetivos, y más considerando la edad de los estudiantes. Por ello, ESSA se ha diseñado para analizar y valorar, exclusivamente, narraciones cortas sencillas.

5. ESTUDIO DE CAMPO

ESSA fue analizado en la tesis doctoral defendida en noviembre de 2009 por el autor del presente artículo: «Análisis del programa informático ESSA, "English Short Story Analyzer", como instrumento para valorar textos escritos en inglés por alumnado español en sexto de Educación Primaria». El estudio consistió en analizar una muestra de 119 textos realizados por alumnado en sexto curso de dos centros públicos de Educación Primaria durante tres cursos académicos. Los estudiantes redactaban en tiempo pasado una historia, cuento o resumen de un libro que previamente habían trabajado en clase. Tenían unos 35 minutos para componer el texto directamente en la pantalla del ordenador, y unos 15 minutos para revisar la ortografía. Las narraciones debían oscilar en torno a las 100 palabras. Esos textos serían después valorados por cuatro docentes y sus calificaciones comparadas con las otorgadas por ESSA.

6. RESULTADOS

Se realizó un análisis de regresión múltiple según el método *stepwise* para determinar la relación entre las variables independientes y la calificación como variable dependiente. El resultado fue un índice de regresión de 0.799 en relación al grado de acuerdo entre la calificación proporcionada por ESSA y la que ofrecieron cuatro docentes evaluando cada uno de los textos de la muestra de acuerdo con una plantilla de valoración. El análisis factorial determinó dos categorías, que se denominaron por las variables que incluían: *gramatical* y *contenido*. El análisis de regresión múltiple fue concluyente al evidenciar que las variables incluidas en el factor gramatical eran las que más contribuían a la hora de predecir la calificación asignada por los docentes. En esa valoración tuvieron especialmente en cuenta la utilización de: verbos y sus tiempos verbales, un vocabulario gramatical básico y las estructuras gramaticales trabajadas en el aula.

7. CONCLUSIÓN

La utilización de ESSA para valorar textos en el área de inglés contribuye a la valoración de la competencia gramatical, ya que las composiciones escritas son un instrumento idóneo, para que el alumnado refleje los aprendizajes adquiridos en un contexto comunicativo significativo y creativo.

Los resultados del estudio de campo han puesto de manifiesto que ESSA realiza el análisis y las valoraciones de los textos con una efectividad similar a la alcanzada por otros programas de evaluación automatizada de textos. El grado de acuerdo entre las puntuaciones de ESSA y las otorgadas por los docentes evaluando los mismos textos es elevada ($r = 0.799$).

Los relatos, por otra parte, se escriben directamente en la pantalla de un ordenador con acceso a Internet, recibiendo el estudiante *feedback* inmediato sobre su trabajo, con la motivación que ello conlleva, como así observaron los docentes que implementaron el programa en la práctica.

ESSA ha sido concebido como un instrumento que ayude al docente en la evaluación de pruebas abiertas y que, además, sea fácilmente integrable dentro de las actividades que los libros de texto proponen para llevar a cabo la clase de lengua extranjera. Su desarrollo supone un paso más a la hora de integrar la tecnología en el campo de la evaluación educativa.

REFERENCIAS BIBLIOGRÁFICAS

Bird, S.; Klein, E. & Loper, E. (2009). *Natural Language Processing with Python*. USA: O'Reilly Media Inc.

Cizek, G. J. y Page, B. A. (2003). The Concept of Reliability in the Context of Automated Essay Scoring, in Shermis, M. D. y Burstein, J. C. *Automated essay Scoring. A Cross-Disciplinary Perspective*. New Jersey: Lawrence Erlbaum Associates, Publishers, 125-145.

Dikli, S. (2006). An Overview of Automated Scoring of Essays. *The Journal of Technology, Learning, and Assessment*, 5(1), 1-35. Disponible en http://escholarship.bc.edu/cgi/viewcontent.cgi?article=1044&context=jtla [Consulta 2007, 10 de febrero]

Elliot, S. (2003). Intellimetric™: From Here to Validity, in Shermis, Mark D. y Burstein, Jill C. *Automated Essay Scoring. A Cross-Disciplinary Perspective*. New Jersey: Lawrence Erlbaum Associates, Publishers, 71-86.

Ericsson, P. F. (2006). The meaning of meaning. Is a paragraph more than an equation? in Ericsson, P. F. y Haswell, R. *Machine Scoring of Students Essays. Truth and consequences*. USA: Utah State University Press, 28-37.

Field, A. (2009). *Discovering Statistics using SPSS*. Third Edition. California: Sage Publications Inc.

Haswell, R. (2006). Automatons and Automated Scoring. Drudges, Black boxes, and Deia Ex Machina, in Ericsson, P. F. y Haswell, R. *Machine Scoring of Students Essays. Truth and consequences*. USA: Utah State University Press, 57-78.

Keith, T. Z. (2003). Validity of Automated Essay Scoring Systems, in Shermis, M. D. y Burstein, J. C. *Automated essay Scoring. A Cross-Disciplinary Perspective*. New Jersey: Lawrence Erlbaum Associates, Publishers, 147-169.

Leacock, C. y Chodorow, M. (2003). Automated Grammatical Error Detection, in Shermis, M. D. y Burstein, J. C. *Automated Essay Scoring. A Cross-Disciplinary Perspective*. New Jersey: Lawrence Erlbaum Associates, Publishers, 195-207.

Myers, M. (2003). What Can Computers and AES Contribute to a K-12 Writing Program?, in Shermis, Mark D. y Burstein, Jill C. *Automated Essay Scoring. A Cross-Disciplinary Perspective*. New Jersey: Lawrence Erlbaum Associates, Publishers, 3-20.

Page, E. B. (2003). Project Essay Grade: PEG, in Shermis, M. D. y Bursten J. *Automated Essay Scoring: A Cross-Disciplinary Perspective*. New Jersey: Lawrence Erlbaum Associates, Publishers, 43-54.

Philips, S. M. (2007). *Automated Essay Scoring: A Literture Review*. Society for the Advancement of Excellence in Education (SAEE), 9-70. Disponible en http://www.saee.ca/pdfs/036.pdf [Consulta 2007, 18 de marzo]

Rudner, L. y Gagne, P. (2001). An Overview of Three Approaches to Scoring Written Essays by Computer. *Practical Assessment, Research & Evaluation. A peer-reviewed electronic journal*, 7(26). Disponible en http://pareonline.net/getvn.asp?v=7&n=26 [Consulta 2008, 26 de enero]

Valenti, S.; Neri, F. Y Ccucciarelli, A. (2003). An Overview of Current Research on Automated Essay Grading. *Journal of Information Technology Ed-*

ucation. DIIGA, 2 319-329. Disponible en http://jite.org/documents/Vol2/v2p319-330-30.pdf [Consulta 2007, 27 de agosto]

Vantage Learning (2006). *Research Summary. IntelliMetricTM Scoring Accuracy Across Genres and Grade Levels*. Disponible en http://www.vantagelearning.com/docs/intellimetric/IM_ReseachSummary_IntelliMetric_Accuracy_Across_Genre_and_Grade_Levels.pdf [Consulta 2009, 29 de mayo]

Vantage Learning (2007). *IntelliMMY Access!® Efficacy Report*. Disponible en http://www.vantagelearning.com/docs/myaccess/myaccess.research.efficacy.report.200709.pdf [Consulta 2009, 29 de mayo]

PALABRAS CLAVE

Evaluación automatizada de textos, evaluación de textos, software evaluativo, evaluación automatizada.

KEYWORDS

Automated essay scoring, writing assessment, computer software assessment, automated evaluation.

PERFIL PROFESIONAL DEL AUTOR

José Luis García González, profesor asociado en la facultad de Educación de la Universidad de Cantabria. Área de Nuevas Tecnologías Aplicadas a la Educación. Doctor en Educación por la UNED. Miembro del grupo de investigación DIM. Las áreas de interés se centran en el procesamiento del lenguaje natural y la recuperación de la información. Más específicamente, en el campo de la evaluación automatizada de textos y la evaluación de respuestas cortas.

Dirección del Autor: José Luis García González
Edificio Interfacultativo. Facultad de Educación.
Avenida de los Castros s/n. Despacho 304. Santander 39005 Cantabria.
Email: garciagjl@unican.es

Fecha Recepción del Artículo: 06. Febrero. 2010
Fecha Revisión del Artículo: 25. Mayo. 2010
Fecha Aceptación del Artículo: 06. Julio. 2010
Fecha de Revisión para publicación: 12. Julio. 2011