Número 1 (2013 Volumen 2) ISSN 0214-9877

DIFICULTADES EN EL ANÁLISIS AUTOMÁTICO DE TEXTOS ESCRITOS

Héctor Díez Caso y Jesús-Nicasio García-Sánchez
Universidad de León
Departamento de Psicología, Sociología y Filosofía, Campus Universitario de Vegazana, s/n, 24071 León.
Email: hdiec@unileon.es, jn.garcia@unileon.es

Fecha de recepción: 23 de septiembre de 2012
Fecha de admisión: 15 de marzo de 2013

RESUMEN

El propósito de este estudio es identificar y analizar las dificultades surgidas durante el proceso de automatización del análisis de textos escritos en base a un protocolo de corrección diseñado para ser aplicado manualmente. Para este estudio se ha contado con una muestra de 237 textos escritos por alumnos con edades comprendidas entre los 8 y los 16 años que han sido analizados tanto manual como automáticamente. Los resultados señalan a los criterios de evaluación poco claros y difícilmente objetivables, como el principal enemigo de los sistemas de análisis automático. Durante la realización de este estudio se recibieron ayudas competitivas del proyecto del MICINN (EDU2010-19250/EDUC, 2010-2013) concedidas al Investigador Principal (J. N. García). Asimismo, H. Diez-Caso ha recibido una subvención de la Consejería de Educación de la Junta de Castilla y León, cofinanciada por el Fondo Social Europeo (EDU/1867/2009, 2009).

Palabras clave: automático de textos, corrección de textos, lingüística computacional, psicolingüística,

ABSTRACT

The purpose of this study is to identify and analyze the difficulties encountered during the process of automating the analysis of written texts based on a correction protocol designed for manual application. For this study a sample of 237 texts, written by students aged between 8 and 16 years, have been analyzed both manually and automatically. The results point to the unclear and difficult to measure evaluation criteria, as the main enemy of automatic analysis systems. During this study we received competitive funds from de MICINN project (EDU2010-19250, 2010-2013) awar-

ded to Principal Researcher (J. N. García). In addition, Héctor Díez Caso has been awarded a grant from the Regional Ministry of Education of the Castilla and Leon Government, jointly funded by the European Social Fund (EDU/1867/2009, 2009).

Key words: Seshat, automatic text analysis, computational linguistics, psycholinguistics, text correction.

INTRODUCCIÓN

La automatización de los procesos de análisis de textos escritos es la aplicación de la ingeniería lingüística a procesos que van desde la mera detección de los errores lingüísticos cometidos en un texto escrito en lenguaje natural (Villena et al., 2002), hasta la interpretación automatizada de complejos protocolos de evaluación psicolingüística (Diez-Caso & García, 2011). En la actualidad, existe un amplio abanico de herramientas disponibles dentro de este espectro de posibilidades. En el ámbito de la revisión lingüística, los ejemplos más conocidos los encontramos integrados en programas de uso diario, desde procesadores de texto hasta navegadores web; no obstante, existen también una serie de aplicaciones de revisión lingüística independientes, igualmente relevantes y dignas de mención, como por ejemplo Stilus (2013) en el caso del español o Textalyser (2013) y la suite WordSmith Tools (2013) en el del inglés. Asimismo, atendiendo a aquellas herramientas que aspiran no sólo a revisar, sino también a evaluar la calidad del texto desde una perspectiva psicolingüística, cabe destacar la existencia de Coh-Metrix (2013) y Seshat (Diez-Caso & García, 2011).

El esfuerzo dedicado a la mejora de la competencia comunicativa escrita en el ámbito escolar resulta a menudo insuficiente en nuestros días (Gilbert & Graham, 2010; Wyse, 2003) y son muchos los docentes que admiten verse desbordados a la hora de afrontar esta parte esencial del proceso educativo (Graham, Gillespie & McKeown, 2013, Kiuaura, Graham, & Hawken, 2009). Afortunadamente, el actual proceso de digitalización de las aulas que se lleva a cabo a través de la incorporación de las nuevas tecnologías en los centros escolares nos brinda la oportunidad de hacer frente a esta problemática a través de herramientas de apoyo como las citadas Coh-Metrix y Seshat, entre otras. No obstante, el desarrollo de este tipo de aplicaciones informáticas resulta a menudo muy complejo debido a las inherentes dificultades que presenta todo proceso de análisis del lenguaje natural que aspire a recrear el proceder de un experto humano entrenado.

El objetivo de este estudio es analizar pormenorizadamente los problemas a los que se enfrenta la automatización del análisis de textos escritos, no desde la perspectiva de las limitaciones impuestas por el procesamiento del lenguaje en sí mismo, sino abordando el problema de recrear la tarea de un experto humano que ampare sus decisiones en un protocolo de corrección previamente establecido. Asimismo, una vez identificadas las dificultades del proceso de análisis automático, se proponen una serie de medidas desarrolladas con el fin de contrarrestar sus efectos.

MÉTODO

Para la realización de este estudio se ha contado con una muestra de 237 textos pertenecientes al corpus lingüístico desarrollado por el Grupo Investigación de Excelencia de la Junta de Castilla y León GR259 haciendo hincapié en la composición escrita como vehículo para la investigación de la escritura en relación a factores de productividad, calidad, estructura y coherencia del texto. Los textos, escritos por alumnos con edades comprendidas entre los 8 y los 16 años, han sido cuidadosamente seleccionados con la intención de crear una muestra lo suficientemente heterogénea como para recrear el mayor número posible de escenarios de trabajo reales a los que podría enfrentarse el sistema de análisis automático en un centro escolar.

Protocolo de corrección

El protocolo de corrección empleado ha sido íntegramente desarrollado por los miembros del equipo de investigación con el fin de establecer unos criterios de corrección estandarizados que permitan evaluar diferentes aspectos de la composición escrita desde una perspectiva psicolingüística (Díez-Caso & García, 2011a, 2011b, García, *et al.,* 2010, García-Martín, & García, 2011). Este protocolo está compuesto por una serie de indicadores y/o medidas que se presentan en dos grandes bloques. El primero es el bloque de las medidas basadas en el texto. Esta medidas se fundamentan en la localización y recopilación de determinadas características de un constructo en el texto, las cuales permiten la verificación textual del constructo que se evalúa. Dentro de esta categoría se evalúan aspectos referidos a la generación de información, productividad y organización de la información. El segundo bloque es el de las medidas basadas en el lector. En este caso, las medidas se basan fundamentalmente en el juicio del lector para determinar si el constructo bajo evaluación está presente o no en el texto, y se evalúan los aspectos referidos a la estructura, coherencia y calidad del mismo. Como puede observarse desde un principio, este segundo bloque es susceptible de una mayor divergencia entre correctores que el primero, dada la naturaleza relativamente subjetiva de los indicadores que lo componen.

Herramienta de análisis automático

Seshat ha sido la herramienta seleccionada para realizar el análisis de los textos. Este software experimental de la Universidad de León que se encuentra actualmente en fase de desarrollo, ha sido concebido para automatizar los procesos de evaluación y análisis psicolingüístico de textos escritos y generar informes exhaustivos de acuerdo a una serie de reglas preestablecidas. En este caso, dichas reglas coinciden con los indicadores presentes en el protocolo de corrección. Asimismo, aunque actualmente su capacidad es limitada, Seshat está ideado para generar estrategias instruccionales adaptadas a cada caso con el fin de permitir a los profesionales de la enseñanza respaldar sus conclusiones en informes previos.

Procedimiento

Con el fin de hallar las limitaciones del análisis automatizado, los 237 textos seleccionados han sido analizados tanto de forma automática como manual; no obstante, este procedimiento de empleo de varios jueces demanda, por su parte, controlar el consenso y acuerdo entre ellos. En el primero de los casos, el consenso entre correctores en este estudio ha venido propiciado por el uso único de un protocolo que incluye normas específicas de corrección de textos y por la formación específica y el entrenamiento que los correctores han recibido previamente, tanto para conocer dicho protocolo, como para aplicarlo en la evaluación real de textos escritos. En el segundo de los casos, para hallar el acuerdo entre correctores en los totales de las medidas textuales obtenidas tras la corrección de los escritos de los alumnos, se ha empleado el coeficiente de Kappa de Cohen. Este coeficiente parece ser el más apropiado en la búsqueda de este tipo de acuerdos, ya que considera y controla el efecto del azar. Dicho coeficiente se ajusta a la siguiente fórmula estadística:

$$k = \sum_{1}^{k} \frac{p_o - p_e}{1 - p_e} \times 100$$

De modo que la probabilidad observada es:

$$p_o = \frac{\text{Acuerdos}}{\text{Acuerdos} + \text{Desacuedos}}$$

Y la probabilidad esperada por azar es:

$$p_{e(ct1,ct2)} = \frac{O_{ct1}}{T} \times \frac{O_{ct2}}{T}$$

Donde "ct1" es el corrector número uno, "ct2" es el corrector número dos, "O" representa las ocurrencias y "T" el total de ocurrencias.

Asimismo, durante el proceso de análisis, especialmente en lo concerniente a la parte subjetiva del protocolo de corrección, se llevaron a cabo sucesivas reuniones donde cada uno de los analistas manifestaba su opinión y ofrecía posibles soluciones a los problemas que hubiesen surgido. Las medidas a tomar para refinar el protocolo de corrección eran discutidas en presencia del ingeniero informático a cargo de la parte técnica del proyecto.

RESULTADOS

En cuanto a la parte objetiva o medidas basadas en el texto se refiere, los algoritmos responsables de la automatización de los indicadores presentes en el protocolo de corrección fueron sometidos a sucesivos procesos de refinamiento, llegando a superarse la barrera del 90% de índice de acuerdo (ver Figura 1). No se observaron por tanto problemas destacables en esta parte al margen de la consabida dificultad que entraña toda incursión en el ámbito del procesamiento del lenguaje natural. A decir verdad, más allá de presentar obstáculo alguno, el extraordinario ahorro de tiempo que ofrece el análisis automático frente al manual, sumado al elevado índice de acuerdo obtenido, lo convierten en una alternativa extremadamente eficiente, pudiendo llegar a analizarse 1000 textos de una media de 200 palabras en poco más de 10s, frente a los 20-30 minutos que se tarda en analizar un único texto de forma manual.

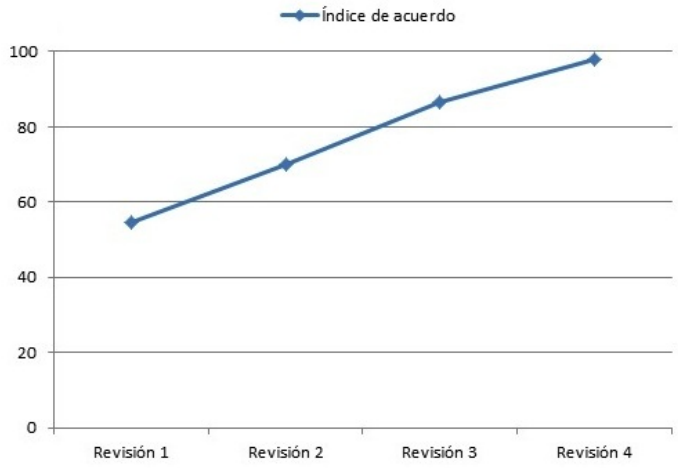

Figura 1. Índice de acuerdo total en productividad

El panorama cambia significativamente cuando hablamos de las medidas basadas en el lector. Obviamente, desde un punto de vista técnico, parece claro que la automatización de este tipo de indicadores acarrearía mayores dificultades dado que, por definición, requieren de la intervención y el juicio de un experto humano, y este tipo de procesos altamente complejos son extremadamente difíciles de recrear. No obstante, a raíz de la comparación de resultados obtenidos por cada uno de los analistas implicados en el proceso de corrección, se constató la imposibilidad de hallar el índi-

ce de acuerdo con respecto a la corrección automática debido a las divergencias existentes entre los propios correctores humanos, llegando incluso esta disparidad, a superar el 40% en algunos casos. Así pues, a la luz de esta nueva información se inició un proceso de revisión del protocolo de corrección intentando aislar las causas de este problema y hallar la mejor manera de afrontarlo. A continuación se detallan los problemas encontrados y las consiguientes dificultades en el proceso de análisis automático.

Medidas difusas: como ya se ha mencionado previamente, uno de los principales objetivos de que cuatro correctores humanos independientes tomasen parte en este estudio era refinar el propio protocolo de corrección gracias sus opiniones y sugerencias, que se pretendía permitiesen objetivar, en la medida de lo posible, aquellos criterios de corrección cuya automatización resultaba más dificultosa. No obstante, las modificaciones sugeridas por los correctores a menudo no se correspondían con las puntuaciones finales. Un análisis más detallado de los resultados obtenidos reveló la causa del problema. La descripción de gran parte de las medidas basadas en el lector resultaba demasiado vaga y era susceptible de múltiples interpretaciones, algo muy común en todos los protocolos de corrección de estas características, donde, para alcanzar el índice de acuerdo deseado entre correctores, se precisa no sólo que dichos correctores conozcan detalladamente del protocolo de corrección, sino que además hayan trabajado juntos en su aplicación.

Criterios de evaluación variables: las calificaciones otorgadas por los correctores humanos tienden a variar incluso en circunstancias similares. Por supuesto, es de esperar que esto suceda, siempre y cuando estas fluctuaciones permanezcan dentro de unos límites aceptables. Desafortunadamente, en el caso del análisis automático este problema cobra una especial importancia, pues dificulta el establecimiento de un criterio general a la hora de valorar los indicadores del protocolo de corrección.

Rangos de calificación desmedidos: a menudo, los protocolos de corrección establecen unos rangos de posibles calificaciones para sus indicadores excesivamente precisos a tenor de los criterios de evaluación de que se dispone. En otras palabras, resulta muy complicado establecer el valor de un indicador de carácter puramente subjetivo llegando incluso a la precisión de los decimales. Obviamente, un análisis automático acarrearía la misma problemática o incluso peor, pues la implementación del sistema resultaría extremadamente compleja al carecerse de criterios de evaluación objetivos.

Repetición de indicadores: en teoría todas las medidas subjetivas pueden analizarse de forma independiente, pero en la práctica resulta difícil hacer distinciones y muchos de estos indicadores se solapan entre sí. Esta situación provoca que las calificaciones arrojadas por los correctores humanos en ocasiones den la impresión de estar sujetas a la mera intención de establecer una diferencia forzosa entre indicadores, dado que no existe un criterio de corrección objetivo y estandarizado que la establezca de por sí. Igualmente complicado resulta automatizar de forma eficiente este tipo de indicadores si no se establecen previamente unas normas que regulen las sinergias existentes entre indicadores.

El número de indicadores en relación al tamaño del texto: teniendo en cuenta que la muestra de estudio la constituyen textos escritos por alumnos de entre 8 y 16 años, resulta indispensable adaptar el protocolo de corrección en función de las dimensiones de la composición escrita, pues la mayor parte de los textos no ofrecen suficiente información para extraer el total de los indicadores presentes en dicho protocolo. Esta limitación afecta tanto al análisis manual como al automático y justifica nuevamente la necesidad de hacer uso de las sinergias existentes entre indicadores.

DISCUSIÓN Y CONCLUSIONES

A la vista de los resultados obtenidos parece indispensable realizar una simplificación del protocolo de corrección que conlleve una drástica reducción del ingente número de indicadores actuales, los cuales, en la mayor parte de los casos o bien están repetidos o bien no aportan datos lo sufi-

cientemente significativos como para constituir una entidad en sí mismos, al menos desde el punto de vista del análisis automático. Una posible solución sería modificar el protocolo de corrección en cuanto a las medidas basadas en el lector se refiere, tomando la calidad final del texto como un indicador general que englobe al resto. Asimismo, los indicadores parciales podrían estar interrelacionados entre sí, pero nunca los indicadores básicos, que mantendrían su independencia en aras de facilitar la automatización. En esta misma línea, antes de iniciar cualquier proceso de análisis automático, el sistema debería tener en consideración tanto el campo semántico como los límites de productividad fijados, en caso de haberlos, o el promedio establecido para el rango de edad en que se encuadre el texto objeto de análisis.

Por otra parte, la evaluación de los textos realizada por expertos humanos se basa en la libre interpretación de unos criterios de corrección a menudo difusos y difícilmente objetivables que terminan por convertirse en unas meras directrices que el corrector conceptualiza en su conjunto y subjetiviza en el momento de su aplicación. Resulta por tanto poco probable que puedan establecerse de criterios de corrección de carácter general susceptibles de ser automatizados. Es de suponer que un experto humano que corrija dos veces el mismo texto, con una diferencia temporal significativa entre ambas correcciones, rara vez lo puntúe igual.

En vista de las dificultades que plantea el análisis automático de textos escritos, una posible solución de carácter general para facilitar esta tarea en lo referente a protocolos de corrección similares al que se ha empleado para llevar a cabo este estudio, pasaría por el establecimiento de una serie de criterios de corrección basados en las posibilidades reales de automatización y en el minucioso estudio de los textos disponibles y de sus correspondientes correcciones manuales. Estos criterios serían de carácter general y se aplicarán a todo texto analizado por igual; no obstante, se ponderarían de diferente manera atendiendo al indicador a obtener y a diversos aspectos coyunturales del texto: edad del autor, límite de productividad, existencia de campo semántico, etc.

REFERENCIAS

Coh-Metrix (2013). *Cohmetrix - Home*. Recuperado de

Díez-Caso, H., & García, J. N. (2011). *Automatización de procesos de evaluación y análisis psicolingüísticos en textos escritos*. Actas VI Congreso Internacional de Psicología y Educación. Valladolid.

Díez-Caso, H., & García, J. N. (2011). *Comparativa de etiquetadores morfológicos para la automatización de análisis textuales*. Actas VI Congreso Internacional de Psicología y Educación. Valladolid.

Díez-Caso, H., & García, J. N. (2011). *Seshat: Análisis Automático de Textos Escritos.* Actas XVIII Congreso Internacional de Psicología Evolutiva y Educativa de la Infancia, Adolescencia, Mayores y Discapacidad: "Desafíos y perspectivas actuales en la psicología". Roma.

García-Martín, E., García, J. N., Pacheco, D. I., & Díez, C. (2009). Design of an Open Corpus and Computer Tool for Writing Development and Instruction among Students 8 to 16 years old, with and without Learning Disabilities. *International Journal of Educational and Developmental Psychology, 21(1), 2,* 107-116.

Gilbert, J., & Graham, S. (2010). Teaching writing to elementary students in grades 4 to 6: A national survey. *Elementary School Journal, 110,* 494–518.

Graham (2013). Writing: importance, development, and instruction. Reading and Writing, 26(1), 1-15. doi: 10.1007/s11145-012-9395-2

Kiuhara, S., Graham, S., & Hawken, L. (2009). Teaching writing to high school students: A national survey. *Journal of Educational Psychology, 101,* 136–160.

Stilus (2013). *Corrector ortográfico, gramatical y de estilo multilingüe | Stilus.* Recuperado de

Textanalyzer (2013). *Text analysis, wordcount, keyword density analyzer, prominence.* Recuperado de

Villena, J., González, B., González, J. C., Muriel, M. (2002). STILUS: sistema de revisión lingüística de textos en castellano. *Procesamiento del lenguaje natural, 29,* 305-306.

Wordsmith Tools (2013). *WordSmith Tools - Mike Scott's Web.* Recuperado de

Wyse, D. (2003). The national literacy strategy: A critical review of empirical evidence. *British Educational Research Journal, 29*, 903–916.

International Journal of Developmental and Educational Psychology
Psicología positiva: desarrollo y educación
INFAD, año XXV
Número 1 (2013 Volumen 2)

ISSN 0214-9877