International Journal of Developmental and Educational Psychology
Desafíos y perspectivas actuales de la psicología
INFAD, año XXIII
Número 1 (2011 Volumen 5)

ISSN 0214-9877

SESHAT: AUTOMATIC ANALYSIS OF WRITTEN TEXTS

Héctor Díez Caso & Jesús Nicasio García
Universidad de León jn.garcia@unileon.es

Fecha de recepción: 3 de enero de 2011
Fecha de admisión: 10 de marzo de 2011

RESUMEN

Este proyecto incide en el uso de las Tecnologías de la Información y la Comunicación (TIC) con la intención de crear una herramienta que automatice el análisis de textos escritos en vistas a generar informes exhaustivos basados en un acervo preestablecido de reglas bien definidas. El software desarrollado, conocido como Seshat, analiza los textos atendiendo a la generación de información, productividad, coherencia y estructura de los mismos. Se intenta, por tanto, suplir la corrección manual que hasta ahora recaía en manos de expertos analistas, y proporcionar una herramienta a psicólogos y docentes que permita ahorrar un tiempo significativo en los análisis textuales, eliminando el error humano del proceso, creando estrategias instruccionales de actuación adecuadas a cada caso y proporcionando a los profesionales la posibilidad de respaldar sus conclusiones en los baremos establecidos a raíz de análisis previos realizados sobre una vasta base de conocimiento que cuenta ya con más de 20.000 textos y que permanece en constante actualización. Durante la realización de este estudio se recibieron ayudas competitivas del proyecto del MICINN (EDU2010-19250) para el trienio 2010-2013; así como del proyecto a Grupos de Excelencia de la JCyL (GR-259; BOCyL 27 de abril de 2009) para el trienio 2009-2011 y con fondos FEDER de la Unión Europea. Ambos concedidos al IP/Director J. N. García.

Palabras clave: Seshat, análisis automático de textos, lingüística computacional psicolingüística, etiquetadores morfológicos

ABSTRACT

This project focuses on developing a software application for automatic text analysis in order to generate comprehensive reports based on a prearranged set of well-defined rules. This application, called Seshat, analyzes texts taking care of information generation, productivity, coherence and structure. It is, therefore, a way to avoid manual analysis and provide teachers and psychologists with a tool that will save time on text analysis and eliminate human error from the equation. On the other hand, this tool is also designed for creating instructional strategies and let professionals support their conclusions on previous reports pulled out from a 20.000 texts database that is continu-

ously updating. During this study we received competitive funds from de MICINN project (EDU2010-19250, 2010-2013); besides from the competitive Project for Excellence Groups JCyL (GR-259; BOCyL 27 on April 2009, 2009-2011) and FEDER funds from de European Union. Both awarded to Principal Researcher (J. N. García).

Key words: Seshat, automatic text analysis, computational linguistics, psycholinguistics, morphological taggers

INTRODUCCIÓN

Seshat es una aplicación informática perteneciente al ámbito de la lingüística computacional desarrollada para automatizar los procesos de evaluación y análisis psicolingüístico de textos escritos con el propósito de generar informes exhaustivos basados en un acervo preestablecido de reglas bien definidas. Esta herramienta se encuadra dentro del marco de las investigaciones llevadas a cabo por el Grupo de Investigación de Excelencia de la Junta de Castilla y León GR259 desde principios de los años 90. Dichas investigaciones, en el ámbito de la escritura, versan en torno a los alumnos con dificultades del aprendizaje, a su pronto diagnóstico y al establecimiento de baremos que permitan identificar perfiles a fin de crear estrategias instruccionales para afrontar este problema. No obstante, la identificación de los problemas en el aprendizaje a través de evaluaciones y análisis psicolingüísticos de textos escritos es un proceso complejo en el que intervienen numerosos factores. Por este motivo, se ha venido desarrollando un protocolo de corrección específico fruto de la experiencia de más de 15 años de investigación (Díez-Caso & García, 2011a, 2011b, García, *et al.,* 2010, García-Martín, & García, 2011). Sin embargo, se ha comprobado que la aplicación manual de estas medidas de corrección resulta a menudo un proceso lento, tedioso y en ocasiones sujeto a la interpretación subjetiva y al error humano. La incorporación de una nueva herramienta, creada al amparo de las Tecnologías de la Información y la Comunicación (TIC), aspira a suponer una importante evolución en el campo de la recién bautizada psicolingüística computacional y más concretamente en la detección y tratamiento de las dificultades del aprendizaje.

Actualmente, existen numerosos estudios que avalan el uso de la computación en el campo de la lingüística funcionalista moderna como medio para el análisis de extensas bibliotecas de textos en lenguaje natural (Atkinson, Ferreira, & Aravena, 2009; Beal, 2009; Biber, 2009; Bird, 2009) o con redes bayesianas; por ejemplo, Yahya, Mahmod, & Ramli (2010). Este campo, denominado lingüística del corpus, tiene uno de sus máximos exponentes en Douglas Biber, autor cuya obra sentó las bases para la creación de una de las primeras herramientas informáticas enfocadas a la identificación y análisis de patrones de uso en textos y su relación con las variables extralingüísticas que pudieran determinarlos, empleando, para ello, técnicas de análisis cuantitativas y cualitativas (Biber, Conrad, & Reppen, 1998). No obstante, desde el punto de vista de la psicolingüística, el análisis de los procesos de la comunicación humana mediante el lenguaje escrito está íntimamente ligado a la tipología del lenguaje empleado y por consiguiente a su estructura y función. La aplicación concebida por Biber ha arrojado resultados interesantes, como en el análisis comparativo de estudiantes con y sin dificultades del aprendizaje de Gregg, Coleman, Stennett, & Davis (2002). Sin embargo, al igual que el resto de las herramientas desarrolladas en este área hasta la fecha, está basada en la lengua inglesa, haciendo imposible su empleo para el análisis de textos en castellano. De este modo, el proyecto Sehsat pretende impulsar el estudio de un área virgen en lengua castellana, permitiendo realizar detallados análisis de la competencia comunicativa escrita de más de 500 millones de hispanohablantes en todo el mundo (Instituto Cervantes, 2007).

Ha sido, por tanto, y continúa siendo el objetivo primordial de este proyecto, el desarrollo, mediante técnicas de inteligencia artificial, de la primera aplicación informática destinada a la auto-

matización de los procesos de análisis, evaluación y corrección de textos escritos en castellano desde una perspectiva psicolingüística. Desde un primer momento se ha pretendido crear una herramienta abierta y flexible, capaz de integrar tantas metodologías de análisis como fuese necesario para adaptarse a las necesidades de uso de diferentes disciplinas; no obstante, en un principio esta aplicación está destinada al campo de la psicología y la educación (Flowerdew, 2009; Xiaofei, 2009), poniendo a disposición del profesorado la tecnología más actual para realizar estudios psicolingüísticos exhaustivos que revelen posibles dificultades del aprendizaje en el ámbito escolar, tanto desde la perspectiva de la evaluación y detección de necesidades instruccionales como desde la enseñanza de la competencia comunicativa escrita.

MÉTODO

Participantes

Se han reunido entre 14000 y 20000 textos escritos de un total de 7232 estudiantes españoles con edades comprendidas entre los 8 y los 16 años, con y sin dificultades para el aprendizaje, que vienen tomando parte en múltiples investigaciones del grupo de excelencia que dirige esta investigación. Todos los textos han sido recogidos por los investigadores del equipo haciendo especial énfasis en la composición escrita como vehículo para la investigación de la escritura atendiendo a factores de productividad, calidad, estructura y coherencia del texto (Arias & García, 2007, de Caso, et al., 2010, Fidalgo & García, 2008, García 2002; García & Arias, 2004; García & de Caso, 2002a; 2004; 2006a, 2006b, 2006c, 2007, García & Fidalgo, 2003; 2004; 2006, 2008a, 2008b; García & Marbán, 2003; García, et al., 2009, Pacheco, García, & Díez, 2009, 2010a; 2010b; Torrance, Fidalgo, & García, 2007).

Previa digitalización de los textos manuscritos, se ha creado un corpus abierto de más de 2 millones de palabras orientado al estudio del desarrollo de la escritura en el ámbito escolar. Dada la ingente cantidad de recursos que supondría llevar a cabo análisis manuales e individualizados de cada uno de los textos, dicho corpus ha sentado las bases para la creación de una herramienta informática destinada a la automatización de los procesos de evaluación y análisis de textos escritos de acuerdo a una serie de medidas, desarrolladas por el grupo de investigación, atendiendo tanto al criterio subjetivo del lector (indicadores globales de calidad, estructura y coherencia), como a los indicadores del propio texto (indicadores objetivos de generación de información o productividad, coherencia y estructura).

Herramientas

El sistema desarrollado consta de tres partes interrelacionadas que han sido abordadas siguiendo un orden lógico. La fase inicial comprendió la selección y adquisición de un etiquetador semántico adaptado a las necesidades específicas del proyecto y su posterior adecuación al sistema desarrollo atendiendo al paradigma de la programación modular para evitar una excesiva dependencia de cualquier software ajeno al proyecto. En cualquier caso, esta decisión se tomó con la intención de obtener resultados en el menor tiempo posible, postergando la implementación de etiquetador comercial propio para más adelante.

Una vez analizadas las diversas opciones disponibles en el mercado en relación a criterios de calidad, facilidad de integración con el resto del sistema, asistencia técnica y coste, finalmente se decidió adquirir el Machinesse Phrase Tagger de la empresa finlandesa Connexor Oy. No obstante, este software no interactúa directamente con el resto de la herramienta sino que lo hace a través de un pseudo-etiquetador del sistema destinado a normalizar las etiquetas empleadas por la herramienta y a corregir los errores surgidos a raíz de las ambigüedades lingüísticas que no hayan sido tenidas en cuenta por el etiquetador comercial.

La segunda fase del desarrollo abordó el diseño e implementación de un analizador de reglas, es decir, la construcción de un sistema experto adaptable a cualquier base de conocimiento expresada en forma de datos y una serie de indicadores interrelacionados susceptibles de ser codificados en forma de patrones de búsqueda. Inicialmente, el diseño se ha centrado en la adaptación y aplicación de las reglas y criterios establecidos por el Grupo de Investigación de Excelencia GR259 para el análisis de composiciones escritas desde el punto de vista del lector y del propio texto, siguiendo los criterios e indicadores que a nivel internacional se están siguiendo en el campo de la escritura, como la red COST European Research Network Learning to Write Effectively (ERN-LWE). Durante todo el proceso de desarrollo y especialmente en este punto se ha hecho hincapié en el empleo de tecnologías fuertemente vinculadas a internet como es el caso de Java (2010) y XML (2010).

La tercera y última fase del proceso de desarrollo de la herramienta se ha centrado en la creación de un interfaz gráfico de usuario que tuviese en cuenta tanto la versión de la herramienta de escritorio, como su vertiente web. La máxima prioridad ha sido diseñar e implementar un entorno intuitivo y de fácil manejo que no requiera de conocimientos previos para su uso. La rapidez de ejecución, eficiencia y sencillez ha primado sobre cualquier otra consideración y se ha hecho uso extensivo de las librerías gráficas de Java, Swing y SWT, para su desarrollo. En síntesis, en esta parte del proceso de desarrollo se ha trabajado codo con codo con psicólogos y docentes recogiendo comentarios y sugerencias para crear un entorno de trabajo innovador, de fácil manejo y que se adapte a sus necesidades reales.

Procedimiento

Los pasos a seguir en esta investigación han comprendido la revisión de antecedentes, la selección del etiquetador semántico comercial, el diseño e implementación de un pseudo-etiquetador del sistema y de un analizador de reglas, así como del interfaz gráfico de usuario y la validación de cada una de las partes de forma individualizada.

Durante la realización del presente proyecto, se ha continuado en todo momento con una revisión exhaustiva, sistemática y permanente de antecedentes específicos del tema y una profundización en el uso de herramientas preexistentes en el campo de la psicolingüística, con objeto de estar al tanto de cualquier detalle relevante para el desarrollo de la nueva herramienta. Se han tenido en cuenta desde analizadores léxicos y morfológicos simples hasta herramientas más avanzadas como la ya mencionada implementada por Biber, Conrad & Reppen (1998) que aún hoy sigue en desarrollo aportando novedades de interés (Biber, 2009).

Procurando siempre la mayor versatilidad, facilidad de uso y difusión posibles, se ha orientado la búsqueda de soluciones técnicas a hacer de Internet una base sólida sobre la que cimentar el desarrollo de la aplicación. De esta forma, partiendo de tecnologías como Java y XML, la generalización de la herramienta vía web una vez terminada la versión de escritorio, ha sido una de las premisas esenciales a tener en cuenta. Igualmente, se ha buscado la solución apropiada para combinar técnicas de inteligencia artificial con esta base tecnológica, intentando mantener un compromiso tanto con su difusión vía web como con su adaptabilidad a diferentes plataformas (Linux, Mac OS, Windows, etc.), atendiendo inicialmente a entornos de desarrollo y ejecución de sistemas expertos como el ya citado CLIPS (2010) y sus variantes como JESS (2010).

Una fase de validación ha seguido a cada etapa del desarrollo, contrastando los resultados obtenidos mediante análisis automáticos con los resultados que se obtuvieron previamente mediante análisis manuales, calibrando, de esta forma, la respuesta automática de acuerdo a las medidas y criterios establecidos con anterioridad por el grupo de investigación. Actualmente se cuenta ya con una versión preliminar de la herramienta capaz de realizar complejos análisis de textos de forma sencilla atendiendo a diversos factores (productividad, calidad, estructura y coherencia) dependien-

tes del nivel, la edad, etc. y ofreciendo estrategias para abordar las diferentes situaciones y/o problemas de forma objetiva.

RESULTADOS

La herramienta informática desarrollada, conocida como Seshat, es una aplicación informática ideada para automatizar los procesos de evaluación y análisis psicolingüísticos de textos escritos y como tal, genera informes de resultados basándose en un acervo preestablecido de reglas bien definidas. Estas reglas responden a un protocolo de corrección desarrollado por el Grupo de Investigación de Excelencia GR259, que contempla dos tipos de medidas de corrección, las basadas en el texto y las basadas en el lector. Dada la complejidad y amplitud de este protocolo de corrección, hasta el momento, la aplicación sólo contempla las medidas basadas en el texto; no obstante, el proyecto Seshat sigue en marcha y se dispone de financiación autonómica y estatal para continuar su desarrollo durante al menos cuatro años más en los proyectos dirigidos por J N García, director del Grupo de Investigación de Excelencia GR259 de la JCyL. La obtención de resultados en el menor tiempo posible ha sido siempre una de las mayores prioridades de este proyecto. De este modo, el diseño del programa ha sido ideado para satisfacer estas premisas iniciales pero sin renunciar a la aplicación del paradigma de la programación modular con el fin de facilitar su mantenimiento en el futuro.

Los resultados de la aplicación, desarrollada de acuerdo al diseño anteriormente expuesto, han superado incluso las expectativas más optimistas. Durante los procesos de verificación formal de Seshat, se ha observado, no sólo un alto índice de coincidencia entre los resultados del análisis automatizado y los previamente obtenidos en análisis manuales, sino también una elevada eficiencia por parte del sistema desarrollado. Así, se ha logrado llegar a analizar más de 1000 textos procedentes de escolares de entre 8 y 16 años en una media de 10 segundos y un corpus profesional con más de 1.500.000 palabras en escasamente 4 minutos y medio.

Hasta el momento, se puede afirmar que la primera fase del desarrollo de este proyecto ha alcanzado los objetivos inicialmente previstos e incluso se podría decir que ha superado ampliamente las expectativas que se tuvieran en un principio.

En la actualidad, Seshat cuenta con un sólido diseño que detalla cada uno de los componentes del sistema a nivel general y, a la luz de este diseño, se han ido implementado aquellas partes que podríamos considerar vitales para el sistema: la integración de un etiquetador comercial como módulo independiente, el desarrollo de un pseudo-etiquetador del sistema que normalice la salida del etiquetador comercial, el desarrollo de un analizador de reglas que aplique al análisis textual las medidas establecidas en el protocolo de corrección desarrollado por el grupo de excelencia GR259 y el posterior desarrollo de una interfaz gráfica de acuerdo a las necesidades del usuario final. En resumen, podría considerarse que la parte construida sienta los cimientos de un proyecto más amplio. No obstante, hay que hacer especial hincapié en que, a día de hoy, la aplicación, aún siendo parte de un proyecto más amplio, ya permite automatizar gran parte de los procesos de evaluación y análisis psicolingüísticos de los textos escritos, con excelentes resultados y el consiguiente ahorro de tiempo por parte de los psicólogos y docentes que ya hacen uso extensivo mismo.

DISCUSIÓN Y CONCLUSIONES

El estudio de corpus lingüísticos ha sido un área de investigación especialmente popular en las últimas décadas dado su potencial de almacenamiento de grandes cantidades de información en formato electrónico; no obstante, la mayor parte de estos estudios no han superado los primeros estadios de su desarrollo, quedándose en lo que podría considerarse como un nivel meramente des-

criptivo. Este proyecto, ha tratado de ir un paso más allá, no sólo aplicando la metodología del corpus a otras áreas de estudio, sino salvando también la distancia existente entre la teoría y su aplicación práctica, de igual forma que viene haciéndose en otras áreas como, por ejemplo, la traducción (Rabadán 2008).

Prácticamente inexistentes son las aplicaciones destinadas al área de la psicología y la educación, salvo contadas excepciones, y sólo en el campo del léxico, no del discurso escrito (Almeda, 2005; Álvarez, Carreiras, & de Vega, 1992a, 1992b; Biber, Conrad, & Reppen, 1998; Justicia, 1995, Justicia, Santiago, Palma, Huertas, & Guitiérrez, 1996). Por esta razón, esta investigación supone un valor añadido a esta disciplina, tanto a nivel teórico como desde el punto de vista de su aplicación práctica. El corpus desarrollado, producto del trabajo llevado a cabo por el equipo de investigación de excelencia GR259 desde hace 16 años, constituye una extensa base de conocimiento que ha sustentado no sólo el estudio empírico de los datos, sino también la creación de una herramienta informática que automatiza los procesos de evaluación y análisis de los textos desde una perspectiva psicológica.

Seshat, destinada al campo de la psicología y la educación, ya es empleada a día de hoy por miembros del grupo de investigación y lo será en un futuro cercano por psicólogos y docentes en colegios e institutos a fin de automatizar la corrección de textos con vistas a evaluar la productividad, calidad, estructura y coherencia de la composición escrita. Entre los usos más frecuentes, también se contempla la valoración del progreso de los estudiantes y la evaluación del nivel de escritura de los alumnos. Hallar los patrones que identifiquen, tanto las dificultades como las habilidades en los textos manuscritos de los estudiantes, es de especial importancia para establecer una referencia documentada sobre alumnos con y sin dificultades del aprendizaje. Dichos patrones cobran también una relevancia especialmente significativa a la hora de asistir a padres y docentes en el proceso educativo, repercutiendo en una indiscutible mejora en el rendimiento académico de los estudiantes sobre todo en la comunicación escrita, algo esencial para la vida diaria.

Esta nueva tecnología también deja el campo abierto para llevar a cabo diversos estudios en el ámbito de las ciencias sociales. Desde el desarrollo de estudios comparativos atendiendo a factores de edad, tipología y nivel de los estudiantes, hasta estudios desde el punto de vista psicosocial y cultural a través de la comparación de los trasfondos y patrones estructurales de las composiciones escritas.

REFERENCIAS

Atkinson, J., Ferreira, A., & Aravena, E. (2009). Discovering implicit intention-level knowledge from natural-language texts. *Knowledge-Based Systems, 22(7),* 502-508.

CLIPS (2010). http://clipsrules.sourceforge.net/, Visited December.

Beal, J. (2009). Creating corpora from spoken legacy materials: variation and change meet corpus linguistics. *Language & Computers, 69(1),* 33-47.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigation language structure and uses*. Cambridge, UK: Cambridge University Press.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics, 14(3),* 275-311.

Bird, S. (2009). Natural Language Processing and Linguistic Fieldwork. *Computational Linguistics, 35(3),* 469-474.

De Caso, A. M., García, J. N., Díez, C., Robledo, P., & Álvarez, M. L. (2010). Enhancing Writing Self-Efficacy Beliefs of Students with Learning Disabilities Improves their Writing Process and Products. *Electronic Journal of Research in Educational Psychology, 8 (1), 195-206.*

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

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

Flowerdew, L. (2009). Applying corpus linguistics to pedagogy: A critical evaluation. International *Journal of Corpus Linguistics, 14(3),* 393-417.

García, J. N., & Arias-Gundín, O. (2004). Intervención en estrategias de revisión del mensaje escrito. *Psicothema, 16 (2),* 194-202.

García, J. N., & de Caso, A. M. (2002). ¿Es posible mejorar la composición en alumnos con dificultades de aprendizaje y/o bajo rendimiento sin que cambie la reflexividad hacia la escritura? *Psicothema, 14 (2),* 456-462.

García, J. N., & de Caso, A. M. (2004). Effects of motivational intervention for improving the writing of children with learning disabilities. *Learning Disabilities Quarterly, 27 (3),* 141-159.

García, J. N., & de Caso, A. M. (2006a). Changes in writing self-efficacy and writing products and process through specific training in the self-efficacy beliefs of students the learning disabled. *Learning Disabilities. A Contemporary Journal, 4 (2),* 1-27.

García, J. N., & de Caso, A. M. (2006b). Comparison of the effects on writing attitudes and writing self-efficacy of three different training programs in students with learning disabilities. International Journal of Educational Research, 43, 272-289.

García, J. N., & de Caso, A. M. (2007). Effectiveness of an Improvement Writing Program According to Students' Reflexivity Levels. *The Spanish Journal of Psychology, 10 (2),* 303-313.

García, J. N., de Caso, A. M., Fidalgo, R., Arias-Gundín, O., & Torrance, M. (2010). Spanish research on writing instruction for students with and without learning disabilities. In C. Bazerman, R. Krut, K. Lunsford, S. McLeod, S. Null, P. Rogers, & A. Stansell (Eds.), *Traditions of Writing Research (pp. 71-81)*. New York & London: Routledge.

García, J. N., & Fidalgo, R. (2003). Desarrollo de la conciencia de los procesos cognitivos de la escritura, mecánicos frente a sustantivos y otros en alumnos de 8 a 16 años. *Psicothema, 15*, 41-48.

García, J. N., & Fidalgo, R. (2008a). Changes in the calibration of writing self-efficacy in students with learning disabilities by gender. *The Spanish Journal of Psychology, 11 (2),* 444-432.

García, J. N., & Fidalgo, R. (2008b). The Orchestration of Writing Processes and Writing Products: A comparison of 6th Grade Students With and Without Learning Disabilities. *Learning Disabilities. A Contemporary journal, 6 (2),* 77-98.

García, J. N., & Fidalgo, R. (2008c). Fostering the self-regulation of the recursive thinking involved in composition writing. En A. Valle, J. C. Núñez, R González-Cabanach, J. A. González-Pienda & S. Rodríguez (Eds.), *Handbook of Instructional Resources & Applications (pp. 171-185).* New York, NY: Nova Science Publishers.

García, J. N., & Marbán J. M. (2003). El proceso de composición escrita en alumnos con DA y/o BR. Estudio instruccional con énfasis en la planificación. Infancia y Aprendizaje. *Journal for the Study of Educational and Development, 26 (1),* 97-113.

García, J. N., Rodríguez, C., Pacheco, D. I., Díez, C. (2009). Influencia del esfuerzo cognitivo y variables relacionadas con el TDAH en el proceso y producto de la composición escrita. Un estudio experimental. *Estudios de Psicología, 30 (1),* 31-50.

García-Martín, E., & García, J. N. (2011, marzo). Corpus usefulness for the learning of writing and development. *Actas VI Congreso Internacional de Psicología y Educación*. Valladolid: COP.

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. *Internacional Journal of Developmental and Educational Psychology, XXI (1), 2*, 107-116.

Gregg, N., Coleman, C., Stennett, R. B., & Davis, M. (2002). Discourse complexity of college writers with and without disabilities: A multidimensional analysis. *Journal of Learning Disabilities, 35 (1*), 23-38, 56.

Instituto Cervantes (2007). http://www.elpais.com/articulo/cultura/espanol/segundo/idioma/estudia/mundo/Instituto/Cervantes/elpepucul/20070426elpepucul_8/Tes, Visited October 2009.

Java (2010). http://java.sun.com/, Visited November.

JESS (2010). http://www.jessrules.com/, Visited December.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus based language studies: and advanced resource book*. Oxon; New York: Routledge.

Pacheco, D. I., García, J. N., & Díez, C. (2010a). Self-regulation of teachers' practice in teaching writing. In J. De la Fuente-Arias and Mourad Ali Essa (Eds.), *International Perspectives on Applying Self-Regulated Learning in Different Settings*. (pp. 553-574). Bern: Peter Lang Publisher.

Pacheco, D. I., García, J. N., & Díez, C. (2010b). Academic performance and the role of self-regulated practice of the teachers in writing. In J. De la Fuente-Arias and Mourad Ali Essa (Eds.), *International Perspectives on Applying Self-Regulated Learning in Different Settings*. (575-594). Bern: Peter Lang Publisher.

Rabadán, R. (2008). 'Refining the Idea of "Applied Extensions" in A. Pym, M. Shlesinger and D. Simeoni (eds) Beyond Descriptive Translation Studies. Investigations in homage to Gideon Toury (pp. 103-117). Amsterdam/Philadelphia: John Benjamins.

Torrance, M., Fidalgo, R., & García, J. N. (2007). The teach ability and effectiveness of strategies for cognitive self-regulation in sixth grade writers. *Learning and Instruction, 17 (3),* 265-285.

Xiaofei, L. (2009). Automatic measurement of syntactic complexity in child language acquisition. International Journal of Corpus Linguistics, 14(1), 3-28.

XML (2010). http://www.w3.org/XML/, Visited December.

Yahya, A., Mahmod, R., & Ramli, A. (2010). Dynamic Bayesian networks and variable length genetic algorithm for designing cue-based model for dialogue act recognition. *Computer Speech & Language, 24(2),* 190-218.