

FIRST: Una herramienta interactiva para el apoyo a la lectura

FIRST: A Flexible Interactive Reading Support Tool

Paloma Moreda, Universidad de Alicante, moreda@dlsi.ua.es

Elena Lloret, Universidad de Alicante, elloret@dlsi.ua.es

Resumen

En nuestra sociedad actual, la información juega un papel fundamental a nivel social, cultural y económico. Aunque esta información no está limitada a Internet, es verdad que la red ha desempeñado un papel fundamental como medio que facilita el acceso e intercambio de información y datos, dando lugar a que la mayoría de la información sea información textual. El reto para las personas es vivir de acuerdo con las exigencias de este nuevo tipo de sociedad, estar informados y actualizados (Wikipedia¹). No tener acceso a la información puede ser una barrera a la hora, por ejemplo, de conseguir un trabajo o comprender las noticias. Entre los grupos de personas que podrían verse afectados por esta situación de exclusión se encuentran las personas con autismo, las cuales suelen presentar dificultades en la comprensión de documentos escritos. Por ello, el proyecto FIRST parte de la necesidad de mejorar la calidad de vida de las personas con autismo y de sus cuidadores, facilitando la lectura de cualquier documento escrito mediante el uso de las tecnologías del lenguaje, las cuales permiten el procesamiento automático de textos escritos, analizando las estructuras que contienen y reduciendo su complejidad y ambigüedad. Como resultado de este proyecto se ha desarrollado una herramienta, denominada Open Book, que adapta los documentos a un contenido más accesible. Por un lado, la herramienta detecta y elimina obstáculos que pueden dificultar la comprensión de un texto (como por ejemplo, palabras ambiguas o poco frecuentes), enriqueciendo el texto original con imágenes, definiciones y/o sinónimos. Por otro lado, la herramienta se puede personalizar y adaptar a las habilidades lectoras de cada persona en particular.

Palabras clave: autismo, tecnología del lenguaje humano, simplificación automática de textos.

Abstract

In our current society, the information plays a key role at different levels (social, cultural and economic). Although the information is not limited to the Internet, the truth is that the Web has become a crucial element to facilitate the information access and interchange, leading to the fact that most of the information available nowadays is textual information. The challenge for people is to live in accordance with the requirements of this new kind of society, be informed and up-to-date (Wikipedia,¹). Not having access to the information can be a barrier to, for instance, get a job or understand the news. Among of groups of people who could be affected by this situation of exclusion are the ones with autism, since they often have difficulty in understanding written documents. Therefore, the FIRST project takes into account the need to improve the quality of life of people with autism and their carers, making easier to read any written document by using human language technologies. These technologies allow the automatic processing of written texts, analyzing their structures and reducing the complexity and ambiguity of the words and sentences contained. As a result of this project, a tool, called Open Book, which adapts the documents to a more accessible content, has been developed. On the one hand, this tool detects and removes obstacles that may difficult the understanding of a text (ambiguous words, rare, highly specialized, expressions of more than one word, metaphors, complex sentences), enriching, at the same

¹ http://es.wikipedia.org/wiki/Sociedad_de_la_informaci%C3%B3n

time, the original text with images, definitions and/or synonyms. On the other hand, this tool can be specifically customized and adapted to the reading skills of each individual.

Keywords: *autism, human language technology, automatic text simplification.*

1. Introducción

En 2011 y bajo la financiación de la Unión Europea y su séptimo programa marco (FP7-2007-2013-nº 287607) un equipo multidisciplinar formado por investigadores en Tecnologías del Lenguaje Humano (TLH) y expertos en trastornos del espectro autista (TEA) se unen con el objetivo de desarrollar una herramienta multilingüe, capaz de (i) identificar obstáculos candidatos a dificultar la comprensión lectora de un documento y (ii) proporcionar información sobre dichos obstáculos, dando lugar a una versión enriquecida del documento. El principal objetivo es mejorar la calidad de vida de las personas con autismo y de sus cuidadores, facilitando la lectura de cualquier documento escrito mediante el uso de las TLH.

¿Por qué autismo? Los TEA son un grupo de discapacidades que afectan al desarrollo neurológico y que se caracterizan por la alteración cualitativa de la comunicación y conductas repetitivas estereotipadas (Mesibov, G.B., Adams, L.W., Klinger, L.G., 1997). Como espectro, estas discapacidades afectan de distinta manera a cada persona, si bien la mayoría presentan dificultad en mayor o menor medida en el correcto entendimiento de los textos (Minschew, N. and Goldstein, G., 1998). Es por ello, que en nuestra sociedad actual, en la que la información juega un papel fundamental a nivel social, cultural y económico, y en la que la cantidad de información disponible en Internet aumenta día a día con un crecimiento exponencial, el acceso a la información se ha convertido en una prioridad. No tener acceso a la información, y que este acceso suponga un privilegio para una parte de la sociedad, puede ser una barrera a la hora, por ejemplo, de conseguir un trabajo o comprender las noticias y puede dar lugar a situaciones de exclusión social.

¿Por qué TLH? Las TLH, y más concretamente, el Procesamiento del Lenguaje Natural (PLN), proporcionan técnicas y herramientas adecuadas para la mejora de la interacción hombre-máquina cuando tratamos con lenguaje natural (Mitkov, 2005). Avances recientes en PLN tales como la resolución de la anáfora (Hendricks, I., Devi, L.S., Branco, A., Mitkov, R., 2011), la desambiguación de las palabras (Navigli, 2009), el reconocimiento de expresiones temporales (Reeves, R.M., Ong, F.R., Matheny, M.E., Denny, J.C., Aronsky, D., Globbel, G.T., Montella, D., Speroff, T., Brown, S.H., 2012) o la detección de lenguaje figurativo/no literal (Li, L. and Sporleder, C., 2010), son ejemplos de procesos que pueden utilizarse para identificar posibles obstáculos en un documento y eliminarlos o sustituirlos por términos o expresiones más fáciles de comprender, o incluso proporcionar información sobre ellos con el fin de facilitar la comprensión de dichos obstáculos.

Por todo ello, parece obvio y necesario que expertos de ambos campos colaboren en la identificación de aquellos obstáculos que más dificultan el correcto entendimiento de los textos a las personas con autismo, y en el análisis de cómo el uso de las TLH podría facilitar la superación de dichos obstáculos.

2. El proyecto FIRST

El proyecto FIRST (<http://www.first-asd.eu>), con una duración de tres años (desde el 1 de octubre de 2011 hasta el 30 de septiembre de 2014), ha sido desarrollado por un equipo multidisciplinar, formado por socios académicos (Universidad de Wolverhampton, Universidad de Jaén, Universidad de Alicante), socios técnicos (Kodar, iWeb Technologies) y socios expertos en autismo (Deletrea, Autism-Europe, Paralell World and Central and North West London NHS Foundation Trust). Además, el entorno multilingüe en el que este proyecto se enmarca posibilita que el trabajo se haya llevado a cabo para textos en Inglés, Español y Búlgaro.

El resultado más destacable obtenido tras el trabajo realizado durante la duración del proyecto, ha sido el desarrollo de un prototipo software disponible *on line*, denominado Open Book (<http://www.openbooktool.net>), el cual adapta de forma automática documentos escritos en cualquiera de las lenguas del proyecto, a un contenido más accesible. Este prototipo se caracteriza por:

- (i) Realizar la identificación de obstáculos sintácticos y semánticos, y su resolución automática cuando la tecnología así lo ha permitido. Mientras que la fase de detección solamente permite a los usuarios conocer si una palabra o expresión podría ser o no un obstáculo, la resolución de los obstáculos detectados se realiza proporcionando elementos de apoyo, tales como definiciones, imágenes, sinónimos o referencias a otros elementos en el texto.
- (ii) Permitir la personalización de las operaciones de simplificación que se realizan en función de las necesidades de cada usuario. Atendiendo a las habilidades lectoras que tenga cada usuario se pueden activar o desactivar funcionalidades asociadas a la detección y resolución de los distintos tipos de obstáculos tratados.
- (iii) Ofrecer dos modos de operación atendiendo a que el usuario sea un cuidador o un usuario final. La principal diferencia entre ambos radica en que el modo cuidador ofrece más posibilidades de revisión, edición y corrección automáticos de los textos con el fin de obtener una versión lo más adaptada posible a las necesidades de cada usuario.

El Cuadro 1 muestra la interfaz de la herramienta OpenBook.



Cuadro 1: Interfaz de Open Book

Por otro lado, y puesto que el objetivo de partida en el proyecto era ayudar a personas con autismo a leer documentos con mayor autonomía, durante el tiempo de vida del mismo se ha evaluado el impacto que una herramienta de estas características tendría sobre las personas con autismo y sus cuidadores. Dicha evaluación ha permitido detectar una mejora en el acceso a la información y por tanto, un incremento en el número de oportunidades de formación profesional, cultural y social, lo que favorece la inclusión social de las personas con autismo.

3. La Universidad de Alicante en el FIRST

Desde el punto de vista científico, el trabajo relativo al análisis y estudio de los recursos lingüísticos que podrían ser utilizados fue abordado por los diferentes socios atendiendo a su experiencia en cada una de las áreas existentes dentro de PLN. De esta manera, la Universidad de Wolverhampton lideró los trabajos relativos a la complejidad estructural de los textos, la Universidad de Jaén los relativos a la personalización del proceso de simplificación y la Universidad de Alicante los relativos a la complejidad semántica de los documentos.

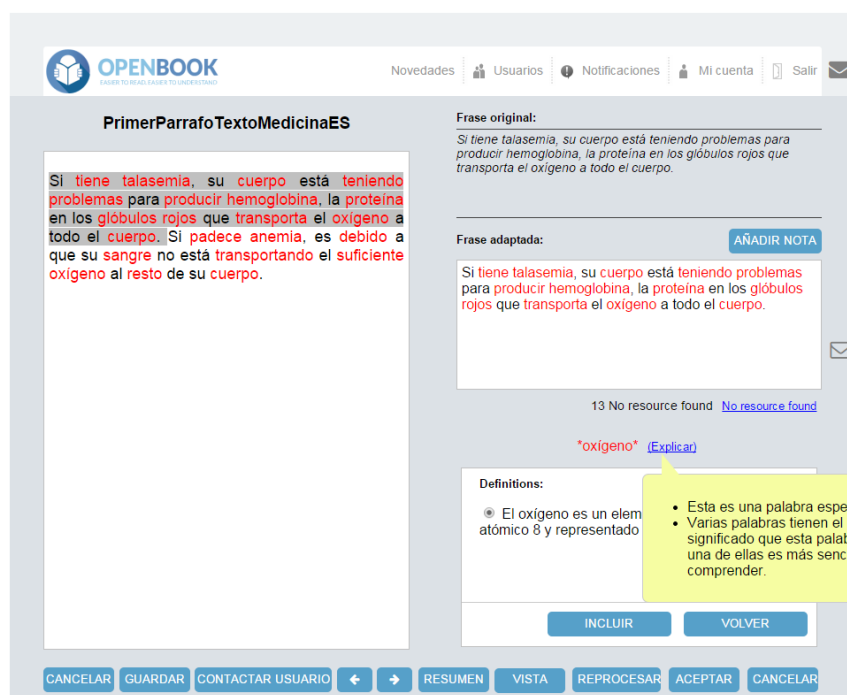
En cada caso, los obstáculos tratados fueron determinados por los resultados obtenidos en el análisis de los requerimientos de usuario llevado a cabo por los socios expertos en autismo. En nuestro caso, dicho estudio estableció que los recursos lingüísticos a tratar debían ser:

- Correferencia, principalmente anáfora pronominal, descripciones definidas y elipsis.
- Lenguaje figurativo.
- Palabras difíciles, incluyendo, palabras polisémicas, verbos mentales, palabras poco frecuentes y, palabras especializadas.
- Expresiones multipalabra.
- Acrónimos y abreviaturas.

Como ya se ha comentado en la sección anterior, el proceso de simplificación y apoyo a la lectura se realiza en dos fases: (i) identificación de los obstáculos en los documentos de entrada, y (ii) resolución de los obstáculos detectados, cuando la tecnología así lo ha permitido, mediante el suministro de elementos de ayuda tales como definiciones, imágenes, sinónimos o referencias a otros elementos del texto procesado. La información que se ha de proporcionar en cada caso fue determinada en el análisis de requerimientos de usuario realizado. De esta manera, para las palabras difíciles la resolución supone la extracción de la definición, una lista de sinónimos de los términos identificados como difíciles y una imagen; para acrónimos y abreviaturas, la resolución es la expansión del término abreviado; para el lenguaje figurativo es una breve explicación de su significado; para la correferencia, el antecedente al que hace referencia; y para las expresiones multipalabra una imagen junto con una definición.

El Cuadro 2 muestra un ejemplo de texto procesado por Open Book cuando es un cuidador el que se valida como usuario. En dicho Cuadro se puede observar el texto procesado a la izquierda, en el que se resaltan los posibles obstáculos en color rojo. La frase considerada en cada momento se marca con fondo gris y a la vez se muestra en el cuadro de la derecha (*frase adaptada*). Además al seleccionar cualquiera de los

términos marcados en rojo, Open Book muestra la información requerida. En este caso se muestra la definición de la palabra *oxígeno* y se explica el motivo por el que esta palabra ha sido seleccionada como posible obstáculo.



The screenshot shows the Open Book interface. On the left, under 'PrimerParrfoTextoMedicinaES', there is a text box with the following text: 'Si tiene talasemia, su cuerpo está teniendo problemas para producir hemoglobina, la proteína en los glóbulos rojos que transporta el oxígeno a todo el cuerpo. Si padece anemia, es debido a que su sangre no está transportando el suficiente oxígeno al resto de su cuerpo.' The words 'talasemia', 'proteína', 'glóbulos rojos', 'transporta', 'oxígeno', 'anemia', 'debido', 'transportando', and 'suficiente' are highlighted in red. On the right, under 'Frase original:', the same text is shown. Below it, under 'Frase adaptada:', the text is: 'Si tiene talasemia, su cuerpo está teniendo problemas para producir hemoglobina, la proteína en los glóbulos rojos que transporta el oxígeno a todo el cuerpo.' Below this, there is a search bar with '*oxígeno*' and a link '(Explicar)'. A definition box is open, showing: 'Definitions: El oxígeno es un elemento atómico 8 y representado'. A yellow tooltip points to the definition, containing the text: 'Esta es una palabra especializada. Varias palabras tienen el mismo significado que esta palabra - tal vez una de ellas es más sencilla de comprender.' At the bottom, there are buttons: 'CANCELAR', 'GUARDAR', 'CONTACTAR USUARIO', 'RESUMEN', 'VISTA', 'REPROCESAR', 'ACEPTAR', and 'CANCELAR'.

Cuadro 2: Ejemplo de texto procesado con Open Book.

Analizamos a continuación con un poco más de detalle cada uno de los recursos lingüísticos abordados.

3.1. Correferencia

La correferencia es un fenómeno que ocurre cuando dos o más expresiones en un mismo texto se refieren a la misma entidad (Radford, 2004). Cuando dos expresiones se correferencian, una es generalmente una forma completa, denominada antecedente, y la otra es una forma abreviada, denominada anáfora. En función del tipo de elemento anafórico se pueden distinguir diferentes clases de correferencia: anáfora pronominal, si el elemento anafórico es un pronombre; descripciones definidas, si el elemento anafórico es un sintagma nominal; o elipsis, si el elemento anafórico se omite.

El siguiente fragmento extraído de uno de los textos con los que se ha trabajado muestra un ejemplo de correferencia. En este texto se puede observar como el pronombre “lo” incluido dentro del verbo “gustar”, corresponde a un elemento anafórico que referencia al antecedente “al café”.

“... la garganta se resistía a dar paso [al café], que tragaba apresuradamente y sin gustar[lo]...”

3.2. Lenguaje figurativo

La detección y resolución de lenguaje figurativo incorpora tres diccionarios, uno para cada lengua tratada en el proyecto. Estos diccionarios fueron construidos a partir de diferentes fuentes de Internet así como de diccionarios de dominio público. Cada entrada en el diccionario contiene una breve definición del elemento descrito. El tamaño de cada uno de los diccionarios obtenidos es de 771 expresiones para el búlgaro, 2.401 para el inglés, y 2.391 para el español.

El proceso de identificación de elementos de lenguaje figurativo en un texto se realiza mediante búsqueda exacta, búsqueda por lema o mediante el uso de expresiones regulares. Por ejemplo, la expresión “*pegarse las sábanas*” sería detectada aún si se encontrara en un documento el texto “*pegar a uno las sábanas*”.

3.3. Palabras difíciles

Dentro de la categoría de palabras difíciles se han englobado:

- *Palabras polisémicas*: palabras para las que existe más de un significado.
- *Verbos mentales* (Grammar and Composition: <http://grammar.about.com/> [last access 20 January 2015]), verbos con un significado relacionado con el entendimiento, descubrimiento, planificación o decisión, como por ejemplo, el verbo descubrir.
- *Palabras poco frecuentes*, aquellas que aparecen con baja frecuencia en los documentos de ámbito general.
- *Vocabulario especializado*, palabras que se utilizan en un dominio específico, como por ejemplo, estomatología, que es la rama de la medicina relacionada con las enfermedades, funciones y estructura de la boca.

La identificación de las palabras difíciles ha necesitado del uso de diferentes recursos en función del tipo de palabra difícil que fuera. Así, para la fase de identificación de las palabras polisémicas se ha hecho uso de Freeling, un analizador morfológico (Padró, 2011). Para los verbos mentales se ha construido un lexicón a partir de información disponible en Internet y de diccionarios de dominio público. La lista definitiva de verbos mentales estaba constituida por 107 verbos para el búlgaro, 159 para el inglés y 74 para el español. La identificación de las palabras poco frecuentes ha hecho uso de las listas de frecuencia de palabras disponibles para cada idioma (Bulgarian Treebank word frequency list², Kucera-Francis word frequency list³, y lista de frecuencias de la Real Academia de la Lengua Española⁴). Por último, la identificación del vocabulario especializado ha accedido a la base de datos léxica WordNet Domains (Bentivogli, L. Forner, P., Magnini, B., Pianta, E., 2004).

² <http://www.bultreebank.org/Resources.html> [last access 20 January 2015]

³ Kucera-Francis wordlist. <http://ota.ahds.ac.uk/headers/0668.xml> [last access 20 January 2015]

⁴ <http://corpus.rae.es/lfrecuencias.html> [last access 20 January 2015]

Para la fase de resolución, en todos los casos la desambiguación se ha realizado utilizando la aproximación del sentido más frecuente, y las definiciones y los sinónimos son extraídos de la base de datos léxica WordNet (Fellbaum, 1998) y de sus versiones para otras lenguas, como Bulgarian WordNet (Koeva, S., Totkov, G., Genov, A., 2004) y MultiWordNet (Atserias, J., Climent, S., Farreres, J., Rigua, G., Rodriguez, H., 1997).

Dado que las definiciones facilitadas por WordNet no siempre resultaron adecuadas para ser comprendidas por personas con TEA, se realizaron estudios sobre si Wikipedia (<http://www.wikipedia.org/> [last access 20 January 2015]) y Wiktionary (<http://www.wiktionary.org/> [last access 20 January 2015]) podrían ser recursos adecuados para proporcionar dicha información. Estos estudios pusieron de relieve, que si bien la información facilitada podría resultar más adecuada para los fines perseguidos, la falta de estandarización de dicha información hacía muy costosa la extracción automática de las definiciones.

3.4. Expresiones multipalabra

Existen sintagmas nominales formados por varias palabras cuyo significado difiere del significado de cada una de las palabras que forman dicha expresión consideradas de forma independiente, por ejemplo, "*cambio climático*". Para la detección de las multipalabras se generó un lexicón formado por aquellas presentes en WordNet y en Wikipedia siempre que el número de enlaces entrantes a esa página fuera igual o superior a 10 (el número de enlaces entrantes a una página representa la importancia de la página). Como resultado se obtuvo un lexicón formado por 1.452.621 palabras para el inglés, 26.288 para el búlgaro y 168.580 para el español. Para su resolución se extrajeron las definiciones de DBPEDIA (Lehmann, j., Isele, r., Jakob, M., Jentsch, A., Kontokostas, D., Mendes, P.N., Hellmann,S., Morsey, M., van Kleef,P., Auer, S., Bizer, C., 2014).

3.5. Acrónimos y abreviaturas

Un acrónimo es una abreviatura formada por las letras iniciales de cada palabra. Por ejemplo, *PLN* es el acrónimo de *Procesamiento del Lenguaje Natural*. Por otro lado, una abreviatura es la forma corta de una palabra, como *Sr*, que es la abreviatura de *señor*. Tanto para la identificación como para la resolución se ha utilizado un lexicón generado a partir de recursos disponibles en Internet y de diccionarios de dominio público. El lexicón resultante consta de 2.160 palabras para el búlgaro, 76 para el inglés y 216 para el español.

4. Conclusiones

Tras los trabajos realizados durante los tres años de duración del proyecto europeo FIRST ha sido posible disponer de un prototipo, denominado Open Book, capaz de identificar de forma automática posibles obstáculos en documentos de entrada y de proporcionar información de apoyo que ayude a la comprensión de dichos obstáculos. Los obstáculos con los que se ha trabajado están relacionados con aquellos que han sido definidos como tales para personas que sufren TEA.

Dentro del trabajo abordado por la Universidad de Alicante en este proyecto, los obstáculos desarrollados han sido la correferencia, el lenguaje figurativo, las palabras difíciles (incluyendo palabras polisémicas, verbos mentales, palabras poco frecuentes y palabras especializadas), expresiones multipalabra y, acrónimos y abreviaturas.

La manera en que este trabajo se ha desarrollado pone de manifiesto cuatro aspectos fundamentales:

- (i) Que la TLH tiene mucho que aportar en el tratamiento, simplificación y enriquecimiento de los textos para que puedan ser accedidos por personas que por algún motivo tengan mermada su comprensión lectora, como es el caso de las personas que sufren TEA.
- (ii) Que el uso de una herramienta como Open Book posibilita la mejora en el acceso a la información y por tanto, un incremento en el número de oportunidades de formación profesional, cultural y social, lo que favorece la inclusión social de las personas con autismo.
- (iii) Que los recursos existentes hasta la fecha no son capaces de proporcionar la calidad, precisión y fiabilidad necesarios para una herramienta como Open Book, por lo que es necesario construir recursos nuevos que se adapten a las necesidades de los usuarios.
- (iv) Que algunas áreas dentro del PLN necesitan madurar para que realmente puedan ser incorporadas a una herramienta de las características de Open Book.

Bibliografía

- Atserias, J., Climent, S., Farreres, J., Rigua, G., Rodriguez, H. (1997). Combining multiple methods for the automatic construction of multilingual WordNets. *Proceedings of the Conference on Recent Advances on NLP (RANLP97)*. TzigovChark, Bulgaria.
- Bentivogli, L. Forner, P., Magnini, B., Pianta, E. (2004). Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. *Workshop on Multilingual Linguistic Resources COLING2004*, (págs. 101-108). Geneva, Switzerland.
- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT.
- (s.f.). *Grammar and Composition*: <http://grammar.about.com/> [last access 20 January 2015].
- Hendricks, I., Devi, L.S., Branco, A., Mitkov, R. (2011). Anaphora Processing and Applications. *8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*. Lecture Notes in Computer Science, Vol. 7099.
- (s.f.). <http://www.wikipedia.org/> [last access 20 January 2015].
- (s.f.). <http://www.wiktionary.org/> [last access 20 January 2015].
- Koeva, S., Totkov, G., Genov, A. (2004). Towards Bulgarian WordNet. *Romanian Journal of Information Science and Technology Vol. 7 No 1-2*, 45-61.
- Lehmann, j., Isele, r., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann,S., Morsey, M., van Kleef,P., Auer, S., Bizer, C. (2014). DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*.

- Li, L. and Sporleder, C. (2010). Linguistic cues for distinguishing literal and non-literal usages. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING'10)* (págs. 683-691). Stroudsburg, PA, USA683-691: Association for Computational Linguistics.
- Mesibov, G.B., Adams, L.W., Klinger, L.G. (1997). *Autism: Understanding the disorder*. New YorkPlenum Press.
- Minschew, N. and Goldstein, G. (1998). Autism as a disorder of complex information processing. *Mental Retardation and Development Disability Research Review* 4, 129-136.
- Mitkov, R. (2005). *The Oxford Handbook of Computational Linguistics (Oxford Handbooks in Linguistics S.)*. Oxford University Press.
- Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys* 41(2).
- Padró, L. (2011). Analizadores multilingües en Freeling. *Linguamática* 3(2), 13-20.
- Radford, A. (2004). *English syntax: An introduction*. Cambridge, UK: Cambridge University Press.
- Reeves, R.M., Ong, F.R., Matheny, M.E., Denny, J.C., Aronsky, D., Globbel, G.T., Montella, D., Speroff, T., Brown, S.H. (2012). Detecting temporal expressions in medical narratives. *International Journal of Medical Informatics*. In press, DOI: 10.1016/j.ijmedinf.2012.04.006.