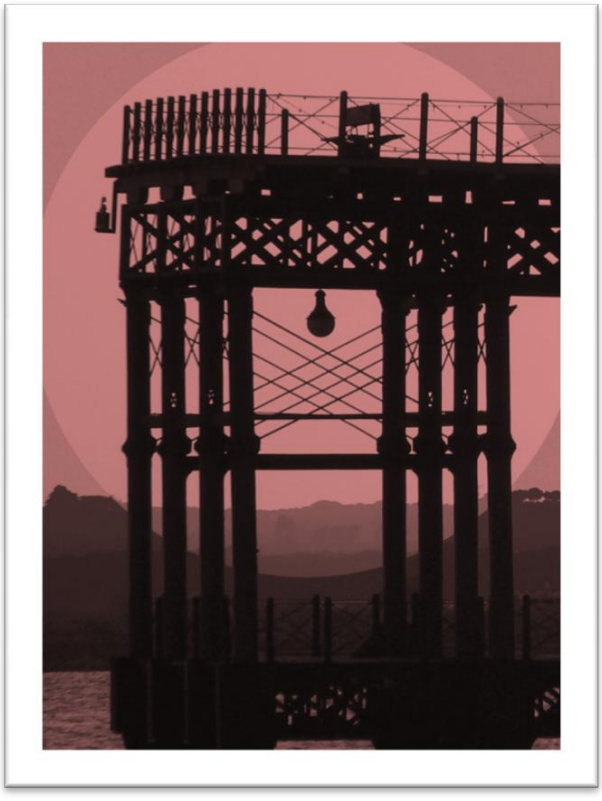
Actas del XXVII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural

Huelva

5 - 7 Septiembre de 2011



Comité de Organización

Presidente:

Manuel J. Maña López

Miembros:

Mariano A. Crespo Azcárate
Noa P. Cruz Díaz
Manuel de la Villa Cordero
Juan Luis Domínguez Olmedo
Jacinto Mata Vázquez
Victoria Pachón Álvarez
Miguel Ángel Vélez Vélez

Comité Científico

Presidente:

Manuel J. Maña López	Universidad de Huelva

Miembros:

José Gabriel Amores	Universidad de Sevilla
Toni Badía	Universitat Pompeu Fabra
Manuel de Buenaga	Universidad Europea de Madrid
Sylviane Cardey-Greenfield	Centre de Recherche Lucien Tesnière. Besançon, France
Irene Castellón	Universitat de Barcelona
Arantza Díaz de Ilarraza	Euskal Herriko Unibertsitatea
Manuel de la Villa Cordero	Universidad de Huelva
Juan Luis Domínguez Olmedo	Universidad de Huelva
Antonio Ferrández	Universitat d'Alacant
Florentino Fernández Riverola	Universidade de Vigo
Mikel Forcada	Universitat d'Alacant
Ana García-Serrano	UNED
Alexander Gelbukh	Instituto Politécnico Nacional, México
Koldo Gojenola	Euskal Herriko Unibertsitatea
Xavier Gómez Guinovart	Universidade de Vigo
José María Gómez Hidalgo	Optenet
Julio Gonzalo	UNED
Ramón López-Cózar Delgado	Universidad de Granada
José Miguel Goñi	Universidad Politécnica de Madrid
Bernardo Magnini	Fondazione Bruno Kessler, Italia
Nuno J. Mamede	Instituto de Eng. de Sistemas e Computadores em Lisboa, Portugal
José Mariño	Universitat Politècnica de Catalunya
M. Antonia Martí	Universitat de Barcelona
María Teresa Martín	Universidad de Jaén
Patricio Martínez-Barco	Universitat d'Alacant
Raquel Martínez	UNED
Paloma Martínez Fernández	Universidad Carlos III de Madrid
Jacinto Mata Vázquez	Universidad de Huelva
Ruslan Mitkov	Universidad de Wolverhampton
Manuel Montes y Gómez	Instituto Nacional de Astrofísica, Óptica y Electrónica, México
Roser Morante	University of Antwerp, Bélgica

Lidia Moreno	Universitat Politècnica de València
Lluís Padró	Universitat Politècnica de Catalunya
Victoria Pachón Álvarez	Universidad de Huelva
Manuel Palomar	Universitat d'Alacant
Ferrán Pla	Universitat Politècnica de València
German Rigau	Euskal Herriko Unibertsitatea
Horacio Rodríguez	Universitat Politècnica de Catalunya
Leonel Ruiz Miyares	Centro de Lingüística Aplicada de Santiago de Cuba
Emilio Sanchís	Universitat Politècnica de València
Kepa Sarasola	Euskal Herriko Unibertsitatea
Isabel Segura Bedmar	Universidad Carlos III de Madrid
Mariona Taulé	Universitat de Barcelona
L. Alfonso Ureña	Universidad de Jaén
Felisa Verdejo Maillo	UNED
Manuel Vilares Ferro	Universidad de A Coruña
Luis Villaseñor-Pineda	Instituto Nacional de Astrofísica, Óptica y Electrónica, México

Revisores adicionales

Emmanuel Aguiano Hernández	Instituto Nacional de Astrofísica, Óptica y Electrónica, México
Alexandra Balahur	Universitat d'Alacant
Zoraida Callejas	Universidad de Granada
Arantza Casillas	Euskal Herriko Unibertsitatea
Alberto Díaz	Universidad Complutense de Madrid
Manuel-Carlos Díaz Galiano	Universidad de Jaén
Miguel A. García Cumbreras	Universidad de Jaén
Meritxell González	Universitat Politècnica de Catalunya
David Griol	Universidad Carlos III de Madrid
Antonio Juárez-González	Instituto Nacional de Astrofísica, Óptica y Electrónica, México
Gorka Labaka Intxauspe	Euskal Herriko Unibertsitatea
Marina Lloberes	Universitat de Barcelona
Héctor Llorens	Universitat d'Alacant
Elena Lloret	Universitat d'Alacant
Montserrat Maritxalar Anglada	Euskal Herriko Unibertsitatea
José Luis Martínez-Fernández	Universidad Carlos III de Madrid
Joaquim Moré	Universitat Oberta de Catalunya
Antoni Oliver González	Universitat Oberta de Catalunya
Jesús Peral	Universitat d'Alacant
José Manuel Perea Ortega	Universidad de Jaén
Juan A. Pérez Ortiz	Universitat d'Alacant
Felipe Sánchez Martínez	Universitat d'Alacant
Esaú Villatoro Tello	Instituto Nacional de Astrofísica, Óptica y Electrónica, México

Preámbulo

Anualmente la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) organiza un Congreso que pretende ser un foro de publicación de artículos científico-técnicos inéditos de calidad relevante en el ámbito del Procesamiento de Lenguaje Natural (PLN) tanto para la comunidad científica nacional e internacional, como para las empresas del sector. El XXVII Congreso de la SEPLN ha sido organizado por el Laboratorio de Recuperación de Información y Minería de Textos y Datos de la Universidad de Huelva y tendrá lugar los días 5 a 7 de septiembre de 2011. El Congreso contará con la presentación de trabajos que incluirán artículos originales, presentaciones de proyectos en marcha y descripciones de herramientas. Así mismo, se llevarán a cabo dos talleres asociados: *ICL: Workshop on Iberian Cross-Language NLP tasks* y *DDIExtraction2011: First Challenge Task: Drug-Drug Interaction Extraction*.

Las áreas temáticas tratadas en el Congreso fueron las siguientes:

- Modelos lingüísticos, matemáticos y psicolingüísticos del lenguaje.
- Lingüística de corpus.
- Desarrollo de recursos y herramientas lingüísticas.
- Gramáticas y formalismos para el análisis morfológico y sintáctico.
- Semántica, pragmática y discurso.
- Lexicografía y terminología computacional
- Resolución de la ambigüedad léxica.
- Aprendizaje automático en PLN.
- Generación textual monolingüe y multilingüe.
- Traducción automática.
- Reconocimiento y síntesis del habla.
- Extracción y recuperación de información monolingüe, multilingüe y multimodal.
- Sistemas de búsqueda de respuestas.
- Análisis automático del contenido textual.
- Resumen automático.
- PLN para la generación de recursos educativos.
- PLN para lenguas con recursos limitados.
- Aplicaciones industriales del PLN.
- Sistemas de diálogo.
- Análisis de sentimientos y opiniones.
- Minería de texto.
- Evaluación de sistemas de PLN.
- Implicación textual y paráfrasis.

Todos los trabajos presentados en el XXVII Congreso de la SEPLN han sido aceptados mediante el proceso de revisión tradicional y ha sido llevado a cabo según el calendario previsto. Queremos agradecer a los miembros del Comité Científico y a los revisores adicionales la labor que han realizado.

En esta edición, se recibieron 78 trabajos, de los cuales 60 eran artículos científicos y 18 correspondían a resúmenes de proyectos de investigación y descripciones de herramientas. De entre los 60 artículos recibidos, 33 fueron finalmente seleccionados para su presentación oral, lo cual fija una tasa de aceptación del 55%. Adicionalmente, se aceptaron otros 12 artículos para su exposición en forma de póster. Autores de otros 10 países han participado en los trabajos presentados en esta edición. Estos países son: Polonia, Bulgaria, Singapur, Portugal, Noruega, Cuba, México, EEUU, Argentina y Brasil.

El Comité Científico del Congreso se ha hecho cargo de la revisión de los trabajos. Este proceso de revisión es de doble anonimato, se mantiene oculta la identidad de los autores que son evaluados y de los revisores que realizan las evaluaciones. En un primer paso cada artículo ha sido examinado de manera ciega o anónima por tres revisores. En un segundo paso, para aquellos artículos que tenían una divergencia mínima de tres puntos (sobre siete) en sus puntuaciones sus tres revisores han reconsiderado su evaluación en conjunto. Finalmente, la evaluación de aquellos artículos que estaban en posición muy cercana a la frontera de aceptación ha sido supervisada por más miembros del Comité.

Estimamos que la calidad de los artículos es alta. El criterio de corte adoptado ha sido la media de las tres calificaciones, siempre y cuando haya sido igual o superior a 5 sobre 7.

Julio de 2011
El Comité de Organización

Artículos **11**

Extracción y Recuperación de Información **13**
Análisis de la expansión de consulta para colecciones médicas utilizando información mutua, ganancia de información y la ontología MeSH
José M. Perea-Ortega, Arturo Montejo-Ráez, Manuel C. Díaz-Galiano y Miguel Á. García-Cumbreras 15
¿De verdad sabes lo que quieres buscar? Expansión guiada visualmente de la cadena de búsqueda usando ontologías y grafos de conceptos
Manuel de la Villa, Sebastián García y Manuel J. Maña 23
Biomedical event extraction using Kybots
Arantza Casillas, Arantza Díaz de Ilarraza, Koldo Gojenola, Maite Oronoz y German Rigau 33
Estudio del uso de Ontologías para la Expansión de Consultas en Recuperación de Imágenes en el Dominio Biomédico
Jacinto Mata, Mariano Crespo y Manuel J. Maña 41
Resolución de Correferencia de Nombres de Persona para Extracción de Información Biográfica
Marcos Garcia y Pablo Gamallo 49
Expansión fonética de la consulta para la recuperación de información en documentos hablados
Alejandro Reyes Barragán, Luis Villaseñor-Pineda y Manuel Montes y Gómez 59
Extracting terminology from Wikipedia
Jorge Vivaldi y Horacio Rodríguez 67
A Spoken Document Retrieval System for TV Broadcast News in Spanish and Basque
Amparo Varona, Silvia Nieto, Luis Javier Rodríguez-Fuentes, Mikel Penagarikano, Germán Bordel y Mireia Diez 77

Resumen Automático **87**
Spanish Text Simplification: An Exploratory Study
Stefan Bott y Horacio Saggion 89
Using Semantic Graphs and Word Sense Disambiguation Techniques to Improve Text Summarization
Laura Plaza y Alberto Díaz 99
COMPENDIUM: Una herramienta de generación de resúmenes modular
Elena Lloret y Manuel Palomar 109

Traducción Automática **119**
Evaluación de estrategias para la traducción automática estadística de chino a castellano con el inglés como lengua pivote
Marta R. Costa-jussà, Carlos Henríquez y Rafael E. Banchs 121

Lexicografía y Terminología Computacionales **129**
Parallel corpus alignment at the document, sentence and vocabulary levels
Rogelio Nazar 131
Unidad discursiva y relaciones retóricas: un estudio acerca de las unidades de discurso en el etiquetado de un corpus en euskera
Mikel Iruskieta, Arantza Diaz de Ilarraza y Mikel Lersundi 139
Extracción automática de colocaciones terminológicas en un corpus extenso de lengua general
Octavio Santana Suárez, José Pérez Aguiar, Isabel Sánchez Berriel y Virginia Gutiérrez Rodríguez 147
AnCora-Net: Integración multilingüe de recursos lingüísticos semánticos
Mariona Taulé, Oriol Borrega y M. Antònia Martí 155

Aprendizaje Automático en PLN **163**
Técnicas de clasificación de opiniones aplicadas a un corpus en español
Eugenio Martínez Cámara, M. Teresa Martín Valdivia, José M. Perea Ortega y L. Alfonso Ureña López 165

Caracterización de Niveles de Informalidad en Textos de la Web 2.0
Alejandro Mosquera y Paloma Moreda 173
Análisis de Sentimientos y Minería de Opiniones: el corpus EmotiBlog
Javi Fernández, Ester Boldrini, José Manuel Gómez y Patricio Martínez-Barco 181
Augmenting Web Page Classifiers with Social Annotations
Arkaitz Zubiaga, Raquel Martínez y Víctor Fresno 191
A Part-of-Speech Tag Clustering for a Word Prediction System in Portuguese Language
Daniel Cruz Cavalieri, Teodiano Freire Bastos Filho, Mário Sarcinelli Filho, Sira Elena Palazuelos Cagigas, Javier Macias-Guarasa y José L. Martín Sánchez 199
Performance analysis of Particle Swarm Optimization applied to unsupervised categorization of short texts
Leticia Cagnina, Diego Ingaramo, Marcelo Errecalde y Paolo Rosso 209
Detección de menciones anidadas basada en expansión para el español
Marcel Puchol-Blasco y Patricio Martínez-Barco 217
Error Analysis for the Improvement of Subject Ellipsis Detection
Luz Rello, Gabriela Ferraro y Alicia Burga 225
Generación automática de reglas de categorización de texto en un método híbrido basado en aprendizaje
Sara Lana-Serrano, Julio Villena-Román, Sonia Collada-Pérez y José Carlos González-Cristóbal 233

Desarrollo de Recursos y Herramientas Lingüísticas 241
ATLAS – Multilingual Language Processing Platform
Maciej Ogrodniczuk y Diman Karagiozov 243
Enriching the Integration of Semantic Resources based on WordNet
Yoan Gutiérrez, Antonio Fernández, Sonia Vázquez y Andrés Montoyo 251
ModeS TimeBank: A Modern Spanish TimeBank Corpus
Marta Guerrero Nieto, Roser Saurí y Miguel Ángel Bernabé Poveda 261
Cognos: A Pragmatic Annotation Toolkit for the Acquisition of Natural Interaction Knowledge
Francisco Javier Calle, Esperanza Albacete, Garazi Olaziregi, Enrique Sánchez, David del Valle, Jessica Rivero y D. Cuadra 271
Handling Reduplication in Basque: A Problem for Spell Checking
Dorota Krajewska y Tamara Hernández Godoy 279
Conversión Fonética Automática con Información Fonológica para el Gallego
Marcos García e Isaac González López 285
Construcción de los WordNets 3.0 para castellano y catalán mediante traducción automática de corpus anotados semánticamente
Antoni Oliver y Salvador Climent 295
Ensinador: corpus-based Portuguese grammar exercises
Alberto Simões y Diana Santos 303

Demostraciones 313

AVI.cat: Asistente virtual pera la mejora de la redacción en catalán
Antoni Oliver, Salvador Climent y Marta Coll-Florit 315
Inferring the Scope of Negation and Speculation Via Dependency Analysis
Miguel Ballesteros, Virginia Francisco, Alberto Díaz, Jesús Herrera y Pablo Gervás 317
MDFaces: An intelligent system to recognize significant terms in texts from different domains using Freebase
Fernando Aparicio, Rafael Muñoz, Manuel de Buenaga y Enrique Puertas 319
MarUja: Prototipo de Asistente Virtual para la Carta de Servicios del Servicio de Informática de la Universidad de Jaén
Eugenio Martínez Cámara, L. Alfonso Ureña López y José M. Perea Ortega 321
Gestión de la información morfológica para la creación de un nuevo par de lenguas con distintos dialectos en un sistema de traducción automática de código abierto
Garbiñe Aranbarri Ariztondo y Itziar Cortés Etxabe 323
Matxin-Informatika: versión del traductor Matxin adaptada al dominio de la informática
Iñaki Alegria, Unai Cabezón, Gorka Labaka, Aingeru Mayor y Kepa Sarasola 325
Sistema de diálogo multimodal basado en modelos estadísticos
E.Sanchis, L.Hurtado, J.A.Gómez, F.García, J.Pastor, J.Planells y E.Segarra 327
Una demostración de Onoma, el conjugador en línea de verbos y neologismos verbales en español
Eduardo Basterrechea, Luz Rello y Rodrigo Alarcón 329
BioViewMed, una herramienta visual de ayuda a la expansión de la cadena de búsqueda usando ontologías
Sebastián García Pérez, Manuel de la Villa y Manuel J. Maña 331
ALICE: Acquisition of Language through an Interactive Comprehension Environment
Maria Fuentes y Meritxell González 333

Proyectos 335

Desarrollo de Recursos para el Análisis Sintáctico Automático del Español: AVALON, una gramática formal y CSA, un corpus sintácticamente analizado
M.ª Paula Santalla del Río 337

Araknion: inducción de modelos lingüísticos a partir de corpora
M. Antònia Martí, Mariona Taulé, Xavier Carreras, Horacio Rodríguez y Patricio Martínez-Barco 339
TextMess 2.0: Las Tecnologías del Lenguaje Humano ante los nuevos retos de la comunicación digital
Patricio Martínez-Barco, M. Antònia Martí, L. Alfonso Ureña y Paolo Rosso 341
Text Simplification in Simplext: Making Text More Accessible
Horacio Saggion, Elena Gómez-Martínez, Esteban Etayo, Alberto Anula y Lorena Bourg 343
AutoIndexer: Investigación y Desarrollo de Metodologías y Recursos Terminológicos de Apoyo para los Procesos de Indexación Automática de Documentos Clínicos
Alberto Díaz, Laura Plaza, Virginia Francisco, Pablo Gervás, Alejandro Palacios, Oliver Partida, Enrique Mota, Arturo Romero e Ignacio Colodrón 345
Procesamiento automático de metáforas con métodos no supervisados
B. Navarro-Colorado, D. Tomás, S. Vázquez, P. Moreda, R. Izquierdo, E. Saquete y F. Llopis 347
MULTIMEDICA: Extracción de información multilingüe en Sanidad y su aplicación a documentación divulgativa y científica
Paloma Martínez, José C. González-Cristóbal y Antonio Moreno Sandoval 349
Spoken language recognition in conversational telephone speech and TV broadcast news (GLOSA)
Luis Javier Rodríguez-Fuentes, Amparo Varona, Mikel Peñagarikano, Mireia Díez y Germán Bordel 351

Pósteres 353

Extracting Information from a Parallel Spanish-English Summary Corpus
Horacio Saggion y Sandra Szasz 355
Domain-neutral, Linguistically-motivated Sentiment Analysis: a performance evaluation
Antonio Moreno-Ortiz, Chantal Pérez Hernández y Rodrigo Hidalgo García 361
Metodología y desarrollo del primer corpus en español anotado con relaciones retóricas
Iria da Cunha, Juan-Manuel Torres-Moreno y Gerardo Sierra 371
Preliminary evaluation of EPEC-RolSem, a Basque corpus labelled at predicate level
Izaskun Aldezabal, María Jesús Aranzabe, Arantza Diaz de Ilarraza y Ainara Estarrona 381
Detección de la polaridad de tweets en español
Eugenio Martínez Cámara, Miguel Ángel García Cumbreras, M. Teresa Martín Valdivia y L. Alfonso Ureña López 391
Georeferencing Textual Annotations and Tagsets with Geographical Knowledge and Language Models
Daniel Ferrés y Horacio Rodríguez 399
Natural Language Processing in Recommender Systems based on Collaborative Filtering
Juan D. Borrero, José Carpio Cañada, Víctor Rivas Santos, Juan J. Merelo Guervós y José L. Álvarez Macías 409
Análisis de preguntas para Búsqueda de Respuestas: evaluación de tres parsers del español
Iria Gayo 419
Minimizando el etiquetado manual en la modelización estadística para la comprensión del habla
Lucía Ortega, Isabel Galiano y Emilio Sanchís 427
Text::Perfide::BookCleaner, a Perl module to clean and normalize plain text books
André Santos y José João Almeida 433
Propuesta metodológica para la creación automática de patrones léxicos usando el Corpus Pattern Analysis
Irene Renau y Rogelio Nazar 443
Detecting source code reuse across programming languages
Enrique Flores, Alberto Barrón-Cedeño, Paolo Rosso y Lidia Moreno 451

Propuesta metodológica para la creación automática de patrones léxicos usando el Corpus Pattern Analysis*

Methodological Proposal for Automatic Creation of Lexical Patterns using Corpus Pattern Analysis

Irene Renau
IULA - Universitat Pompeu Fabra
C/Roc Boronat, 138, 08018 Barcelona
irene.renau@upf.edu

Rogelio Nazar
IULA - Universitat Pompeu Fabra
C/Roc Boronat, 138, 08018 Barcelona
rogelio.nazar@upf.edu

Resumen: El presente artículo propone una metodología para la creación automática de patrones léxicos basada en el Corpus Pattern Analysis (CPA) de Hanks. El CPA se fundamenta en la idea de que las palabras no tienen significado en sí mismas, sino que lo adquieren en contexto. Se aborda la cuestión desde las necesidades de la lexicografía, especialmente de aprendizaje, que demanda cada vez más la incorporación de información gramatical y acorde con la lengua actual en uso. Se ejemplifica dicho procedimiento con el verbo *volcar*.
Palabras clave: Corpus Pattern Analysis (CPA), desambiguación semántica, lexicografía computacional, lingüística de corpus

Abstract: This article presents a methodology for the automatic creation of lexical patterns based on Hank's Corpus Pattern Analysis (CPA). CPA is a corpus analysis method based on the idea that words do not have meaning in isolation, but they acquire it in context. We tackle the problem from the needs of lexicography, especially the pegadogical lexicography, which is demanding more grammatical information, appropriate to language in use. We exemplify the procedure with the verb *volcar* (= 'knock over').
Keywords: computational lexicography, corpus linguistics, Corpus Pattern Analysis (CPA), word sense disambiguation

1. Introducción y objetivos

Este trabajo tiene como objetivo presentar una metodología de detección y configuración automáticas de patrones léxicos basada en el Corpus Pattern Analysis, CPA (Hanks, 2004; Hanks, En prensa), enfocada especialmente desde las necesidades de la lexicografía. El CPA es un método basado en corpus para la detección de los patrones normales de uso de una palabra, en relación con sus rasgos sintácticos y semánticos. A diferencia de las aproximaciones tradicionales basadas fuertemente en la noción de las acepciones vinculadas a los distintos sentidos que una palabra puede tener, el CPA clasifica los distintos usos de la palabra en función de una serie limitada de dichos patrones extraídos de los contextos de uso. El objetivo del presente trabajo es aplicar el procedimiento propuesto por Hanks al análisis de verbos en castellano y explorar las posibilidades de automatización de esta tarea. En concreto, la metodología que se expondrá a continuación surge de una necesidad tanto general, en el panorama lexicográfico, como particular, en relación con un proyecto de *Diccionario de aprendizaje de español como lengua extranjera* (DAELE) que se encuentra en desarrollo (Battaner y Renau, 2008), y del cual se ofrece en línea un primer núcleo de verbos (http://www.iula.upf.edu/rec/daele). Por otro lado, el propósito del trabajo se vincula no solamente con la actividad previa a la publicación de los datos, sino también con

* Este trabajo ha recibido una subvención de los siguientes proyectos del MCI: "Agrupación semántica y relaciones lexicológicas en el diccionario", dir.ª: J. DeCesaris (HUM2009-07588/FILO); APLE: "Procesos de actualización del léxico del español a partir de la prensa", período 2010-2012, dir.ª: M. T. Cabré (FFI2009-12188-C05-01/FILO). También ha recibido subvención de la Fundación Comillas en relación con el proyecto DAELE.

la traslación de estos a la interfaz de usuario de forma dinámica y acorde con la tecnología actual.

El artículo seguirá la siguiente estructura: en primer lugar, se indicarán las necesidades de la lexicografía, en especial de aprendizaje, en relación con la automatización de tareas, como justificación del procedimiento que se describirá; en segundo lugar, se detallará el CPA como método de análisis de corpus; en tercer lugar, se explicará el procedimiento de clasificación y generación de patrones léxicos, ejemplificándolo con el verbo *volcar*; por último, se abordarán las perspectivas de trabajo futuras en relación con la implementación de esta metodología.

2. *Lexicografía de aprendizaje y procesamiento del lenguaje natural*

Desde que Sinclair inauguró la era de los diccionarios basados en corpus con el *Collins Cobuild English Language Dictionary* (J. Sinclair, 1987), el análisis de corpus constituye el procedimiento estándar admisible para la elaboración rigurosa de diccionarios y cualquier otra herramienta lexicográfica. En la actualidad, la necesidad de dotar de mayor rigor la explotación de corpus, así como de agilizarla y facilitar la extracción de información, constituyen necesidades fundamentales para los equipos lexicográficos tanto investigadores como comerciales. Así pues, se trata de alcanzar objetivos de mayor rigor metodológico y teórico, pero también mayor competitividad y rendimiento en cuanto al tiempo y al esfuerzo dedicados. Apresjan (2002) define la "lexicografía sistemática" como una reconstrucción de los significados abstractos basada en los patrones de uso gramaticales y semánticos, y aboga por un "integrated dictionary" que reúna no solo semántica sino también gramática.

En este sentido, se ha demandado a menudo, sobre todo en relación con los *learner's dictionaries* (inaugurados con el citado diccionario Cobuild) incorporar más información gramatical en la microestructura de la entrada lexicográfica (Boogards, 2010; Bosque, 2006). Los diccionarios ya no se conciben solo como repositorios de información semántica con pequeñas indicaciones sintácticas suplementarias (como la categoría gramatical), sino que se ven como repertorios alfabéticos en los cuales un usuario puede consultar las dudas puntuales que tenga sobre cierta palabra, sean estas dudas semánticas o sintácticas. En este panorama se hacen aún más necesarias las herramientas que permitan detectar esta información gramatical e incorporarla de forma sistemática al diccionario.

Aparte de las necesidades de fundamentación teórica y metodológica, de ahorro de costos y de mayor explotación de la información gramatical, un cuarto objetivo justifica que el análisis de corpus se ejecute de forma cada vez más refinada y con más prestaciones: el paso de los diccionarios en papel a la publicación en la web. Pese a que Internet se popularizó en los años noventa, los diccionarios en línea existentes hoy en día no cubren todas las prestaciones que la tecnología les permite ofrecer, de modo que son muy pocos los diccionarios que incorporan elementos complementarios a la información clásica que se ofrecía ya en papel. A causa de ello, los especialistas en español como lengua extranjera y los metalexicógrafos demandan desde hace tiempo la incorporación en el diccionario de hiperenlaces, abundancia de información, múltiples opciones de búsqueda y otros elementos relacionados con la interacción usuario-web y con el hecho de que el espacio en Internet ya ha dejado de ser un problema, antes endémico de los diccionarios.

El análisis de corpus es la actividad que más tiempo consume de las que se requieren en la elaboración de un producto lexicográfico. Así, esta metodología, que se prevé implementar en el corto plazo, pretende no solamente cubrir necesidades teóricas y metodológicas que justifiquen el procedimiento empleado, sino que también pretende cubrir necesidades de coste económico y en relación con los extensos calendarios a los que un proyecto de diccionario se somete. Las relaciones entre la lexicografía y el procesamiento del lenguaje natural han sido ya explicadas (Wilks, Slator, y Guthrie, 1996; Corréard, 2002; Fontenelle, 2002). Se han realizado avances en la incorporación de herramientas computacionales en lexicografía, sobre todo encaminadas a la mejor explotación del corpus, por ejemplo, en relación con la detección de ciertos colocados o de ciertas estructuras gramaticales: Sketh Engine, de Kilgarriff et al. (2004), o DANTE Database, de Convery et al. (2010), son una muestra; pero la vinculación directa entre el corpus y el

diccionario (mediada por la supervisión humana experta) está aún en su fase inicial. Así pues, parece necesario ofrecer alternativas computacionales a la tarea costosa en tiempo y esfuerzo del análisis manual del corpus y de la incorporación de los datos tanto semánticos como sintácticos al diccionario.

3. Metodología de análisis de corpus: el CPA

Como se ha avanzado en la introducción, los principales objetivos del CPA son identificar patrones normales de uso de las palabras y asociarlos con un significado, que no se contempla aisladamente, sino en su contexto de uso. El CPA se fundamenta en la Theory of Norms and Exploitations, TNE (Hanks, 2004; Hanks, En prensa), según la cual el comportamiento sintáctico-semántico de una unidad léxica se establece mediante normas (o usos más frecuentes) y sus desviaciones o "explotaciones", que son usos específicos y no estandarizados que se emplean en contextos poco frecuentes. Por ejemplo, la frase "Le acusaban de desórdenes públicos por haber volcado un contenedor" se considera un uso normal del verbo *volcar* (v. apartado 5), pero la frase "El verano volcaba ya oleadas de turismo" se considera una explotación en la que un sujeto inanimado e inmaterial (una estación del año) ejecuta una acción. En este caso, se está explotando uno de los patrones normales de uso del verbo para crear una metáfora en la que probablemente se personifica al verano. La TNE se basa en la teoría de prototipos (Pustejovsky, 1995), que postula una relación jerárquica entre tipos semánticos (o significados abstractos); dicha concepción se unifica en la TNE con la de *lexical item* de Sinclair (1998), según la cual el significado se asocia fuertemente a un contexto. En última instancia, la distinción nos remite a la obra de Hjelmslev (1943), concretamente a la distinción hecha por él entre semióticas denotativas, en las que un plano de la expresión está en función de un plano del contenido, y las semióticas connotativas, en las que ambos planos forman un nuevo plano de expresión para un nuevo plano del contenido.

El CPA puede considerarse una de las posibles aplicaciones de la TNE, focalizada en la investigación lexicológica y aplicable a la lexicografía. Este modelo de análisis de corpus se está desarrollando para el inglés a través del *Pattern Dictionary of English Verbs*, PDEV (Hanks y Pustejovsky, 2000; Cinková, Holub, y Smejkalová, 2010), y se está comenzando a implementar para el castellano (Renau y Alonso, En preparación) y el italiano (Jezek y Frontini, 2010). Existen unos 700 verbos ya redactados en el PDEV, y se han redactado unos 120 de la versión española. El PDEV es de acceso libre (http://deb.fi.muni.cz/pdev).

Se considera que el CPA resulta útil para fundamentar en él la tarea de análisis de corpus en relación con un proyecto lexicográfico porque cada patrón puede asociarse con la correspondiente acepción del artículo (al menos en el caso de los verbos, que es lo que se ha probado tanto en el PDEV como en los CPA italiano y español). Así pues, constituye un modo de sistematizar la microestructura de la entrada lexicográfica, y permite evitar la subjetividad del lexicógrafo en el análisis de corpus, un problema que a menudo ha sido relacionado con los diccionarios (Atkins y Rundell, 2008).

El CPA permite, en resumen, reunir en un patrón de uso las características sintácticas del verbo, señalar los argumentos y relacionar estos con una ontología que se encuentra en proceso de elaboración (pues se establece *bottom-up*, en vez de *top-down*), que sirve para etiquetar cada argumento mediante un tipo semántico. Es el cruce entre el análisis sintáctico y semántico lo que hace muy preciso y completo el CPA, pero particularmente el hecho de que este análisis semántico cuente con una ontología que se suma a la identificación de argumentos.

4. Propuesta de detección automática de patrones léxicos en el corpus mediante CPA

En relación con las necesidades de más rigor y rapidez en la extracción de información del corpus en relación con el proyecto DAELE, nos proponemos una doble tarea. Por un lado, perseguimos el objetivo de clasificar automáticamente cada uno de los contextos reales de uso del verbo (obtenidos de corpus de prensa en castellano) según los patrones que hemos encontrado para este verbo. Por otro lado, pretendemos generar automáticamente los distintos patrones que presenta el verbo. La metodología que proponemos está basada en el supuesto de que disponemos de una taxonomía del castellano y un conocimiento de la sintaxis de los grupos

nominales en castellano; así, la primera tarea consiste en crear dicha taxonomía (Nazar y Janssen, 2010; Nazar, Vivaldi, y Wanner, En prensa). Para cada contexto verbal analizado, nuestro algoritmo identificará aquellos grupos nominales que aparecen como argumentos del verbo. Acto seguido, estos grupos nominales serán buscados en la taxonomía para reemplazarlos por los hiperónimos a los que aparecen asociados, continuando el ascenso por la taxonomía hasta llegar a alguna de las denominaciones altamente abstractas que aparecen en la ontología de Hanks.

Es relevante para el experimento el hecho de que lo que se pretende no es (o no es prioritariamente) etiquetar correctamente la mayoría de concordancias, sino encontrar los principales patrones de uso de cada palabra. Por ejemplo, si en una muestra de 1.000 concordancias hay un 75 % de ellas mal anotadas, pero el 25 % restante ha ofrecido los patrones léxicos correctos de la palabra analizada, el resultado puede considerarse positivo, porque se persigue principalmente hallar los patrones de uso normales de cada unidad, no un porcentaje elevado de éxito en la adjudicación de patrones a cada concordancia.

La propuesta que se presenta consiste en reflejar en procedimientos automáticos tareas que hasta ahora se realizaban manualmente, y añadir otras que el análisis manual no permite. Se observan las siguientes necesidades, que se corresponden con las distintas fases de análisis de corpus y la creación de patrones: 1) Detección de los argumentos del verbo; 2) Asociación de los argumentos a los tipos semánticos de una ontología; 3) Sustitución de los argumentos por los tipos semánticos para crear el patrón léxico; 4) Introducción de los patrones en la base de datos lexicográfica, como base para la redacción de la entrada; y 5) Asociación de cada acepción del diccionario con los patrones detectados y con las concordancias de corpus vinculadas a cada patrón.

Se exponen a continuación estos pasos del procedimiento.

4.1. Detección de los argumentos del verbo

La detección de los argumentos del verbo en el corpus se proyecta, por el momento, como dependiente de lengua y se configurará teniendo en cuenta los supuestos teóricos básicos acerca del concepto de argumento y de las funciones sintácticas (RAE, 2009). Se tomará como punto de partida la herramienta de análisis lingüístico Freeling (Carreras et al., 2004), que tiene la ventaja de haber desarrollado un analizador de dependencias y ser de código libre. La mayor complejidad en la detección de argumentos del verbo radica en distinguir argumentos de adjuntos, lo cual ocurre en el caso de los complementos de régimen frente a los circunstanciales, por ejemplo. Sin embargo, es más fácil detectar argumentos que se encuentran en posición de sujeto, objeto directo o indirecto: solamente dicha detección es ya un avance para acelerar y hacer más riguroso el análisis de corpus con vistas a su explotación lexicográfica.

4.2. Asociación de los argumentos a los tipos semánticos de una ontología

Todo diccionario podría definirse como una ontología o taxonomía "plana", expuesta de forma lineal, en la que sus lemas se interconectan entre sí mediante una red de relaciones hiperónimo-hipónimo que crea agrupaciones semánticas. Sin embargo, es bien conocido el problema de los círculos viciosos, pistas perdidas y otros defectos de este sistema en los diccionarios hasta hace muy poco realizados sin más herramientas que la pericia de cada lexicógrafo. El trabajo lexicográfico sustentado en una ontología ayuda a sistematizar estas relaciones, y a explotar las posibilidades de ofrecer la información de modo diferente al tradicional, a través de redes relacionales entre las palabras de la macroestructura, más allá de sinonimia, antonimia y otras clásicas relaciones.

La ontología de Hanks, que se va elaborando conforme se van analizando verbos, se establece, como cualquier ontología, mediante tipos semánticos conectados jerárquicamente. Por ejemplo, el tipo semántico "Event" se subdivide en "Process" y "Activity": "Process" se refiere a los eventos no controlados por el ser humano (como enfermedades o fenómenos atmosféricos), y "Activity" denota las acciones humanas.

El empleo de la ontología, clave para la propuesta metodológica expuesta, se observa como una paradoja: por un lado, utilizar la ontología de Hanks en su estadio actual puede ser un primer paso, pero no permitiría etiquetar todos los argumentos con su tipo semántico, porque dicha ontología está avan-

zada pero incompleta; por otro lado, emplear otra ontología ya terminada implicaría incorporar al sistema una herramienta no optimizada para este y requeriría adaptarla. El propio Hanks ha rechazado el uso para CPA de ontologías como WordNet (Miller et al., 1988) porque contemplan las palabras aisladamente, sin tener en cuenta el contexto de uso (Hanks, En prensa). Uno de los objetivos principales del CPA es precisamente vincular los diversos significados de las palabras a sus respectivos contextos.

En cuanto a las explotaciones, se prevé que sean numerosas, pero muy diversas, de modo que no puedan acumularse en patrones unificados (pues si lo hiciesen dejarían de ser explotaciones para ser usos normales). Así, las concordancias que no pudieran agruparse bajo patrones de frecuencia alta serían colocadas en un grupo aparte y clasificadas como explotaciones.

4.3. Sustitución de los argumentos por los tipos semánticos y creación del patrón léxico

Esta fase consiste en tomar el verbo buscado en el corpus y sustituir sus argumentos (detectados en la fase previa) por aquellos que aparecen en la ontología como elementos semánticos hiperónimos de los hallados en la concordancia. El límite de abstracción del tipo semántico de la ontología seleccionado para ser incorporado al patrón se corresponde con aquel que agrupa en niveles inferiores de la jerarquía un gran número de tipos semánticos. Los patrones resultantes deben tener las mismas características que tienen los patrones realizados manualmente, y deben respetar los códigos tipográficos y la metalengua. Igualmente, deben ordenarse por frecuencia decreciente y deben ofrecerse los porcentajes de frecuencia de cada uno (dato que podría mostrarse también en el propio diccionario).

4.4. Introducción de los patrones en la base de datos lexicográfica

Los patrones creados para cada verbo se exportan a la base de datos, como uno de los componentes del campo acepción. Es decir, a cada acepción le corresponde un patrón de CPA, y dicho patrón es utilizado por el redactor como base para la redacción de la definición. La adaptación del patrón a las necesidades del usuario es ya una tarea propiamente lexicográfica.

No está previsto que el patrón sea visible para el usuario, pues se considera que un diccionario didáctico no debe ofrecer información tan técnica a un aprendiz; en cambio, los patrones sí se consideran clave para guiar al lexicógrafo en la configuración tanto sintáctica como semántica de cada acepción (como ya se ha advertido, apartado 4, inicio). La definición natural, que han adoptado la serie Cobuild y otros diccionarios didácticos (entre ellos el DAELE), puede entenderse como una paráfrasis del patrón explicada de modo pedagógico y enriquecida semánticamente.

4.5. Asociación de las concordancias de cada patrón con su correspondiente acepción

Por último, cada definición elaborada con base en su respectivo patrón se asocia con las concordancias de corpus, etiquetadas con dicho número de patrón, con el fin de que sean empleadas como ejemplos. Así, estas concordancias se exportan a la base de datos vinculándolas a su acepción, del mismo modo que se ha hecho con el patrón (apartado anterior). Con vistas a la adaptación de las concordancias a los criterios lexicográficos (relacionados con las necesidades y características del usuario), parece requerirse que estas puedan ser modificadas (extractadas, corregidas ortográficamente, etc.), pese a que la meta para el diccionario que se proyecta es que no se modifiquen excesivamente (en la línea de los Cobuild proyectada por Sinclair). Igualmente, se considera apropiado que dichos ejemplos puedan ordenarse y que puedan marcarse aquellos considerados más relevantes desde el criterio lexicográfico, con respecto a los menos apropiados.

5. Un ejemplo: el verbo "volcar"

Ejemplificaremos el procedimiento mediante el verbo *volcar*. En castellano, este verbo puede utilizarse de distintas formas, por ejemplo, un vehículo puede volcar, pero también una persona puede volcar un líquido sobre una superficie. Claramente se trata de distintos usos, pero en la lexicografía de Hanks, en lugar de representar distintas acepciones, estos usos se representan en patrones. En el verbo *volcar* podemos encontrar, por ejemplo, los siguientes cuatro pa-

trones (aunque en realidad hay por lo menos ocho); se añade un ejemplo de corpus para cada uno de ellos:

- Patrón 1: [[Vehicle]] volcar

Implicatura: [[Vehicle]] turns over, capsizes

Ejemplo: "Otro autobús volcaba en un lugar próximo a la presa de Chira".

- Patrón 2: [[Animate - Process]] volcar [[Artifact]]

Implicatura: [[Human - Process]] knocks over [[Artifact]]

Ejemplo: "Le acusaban de desórdenes públicos por haber volcado un contenedor".

- Patrón 3: [[Human]] volcarse (en - con [[Activity]])

Implicatura: [[Human]] dedicates lot of efforts and enthusiasm to make an [[Activity]]

Ejemplo: "Hay que volcarse en los problemas sociales".

- Patrón 4: [[Human 1]] volcarse [[con Human 2]]

Implicatura: [[Human 1]] dedicates lot of attention to [[Human 2]] or enthusiastically encourages him/her

Ejemplo: "...una afición que ayer se volcó con su equipo".

El verbo *volcar* aparece en 942 concordancias del corpus empleado actualmente para el desarrollo del CPA en castellano, del cual están tomados los ejemplos anteriores. Se trata de un corpus de prensa de 50 millones de palabras. De estas concordancias, se analizaron manualmente 250, que se dividieron en los patrones mencionados, descritos mediante las correspondientes "implicaturas" o significados potenciales que deben inferirse en cada caso. Las etiquetas entre corchetes corresponden a los tipos semánticos de la ontología de CPA, el objeto de la cual es hacer abstracción sobre el tipo de entidad que aparece como argumento del verbo analizado. El sentido de estos patrones es expresar que, en castellano, uno de los usos de *volcar* puede ser el accidente de un vehículo, en el que el vehículo aparece como el único argumento del verbo. Pero también, según el patrón 2, un agente puede volcar un artefacto, tal como un manifestante puede volcar un contenedor. En un uso totalmente distinto, según el patrón 3, una persona puede volcarse en una actividad o bien, en otro uso distinto del verbo, según el patrón 4, una persona puede volcar una sustancia sobre una superficie.

Otros usos de *volcar* hallados en el corpus se dibujan de forma menos precisa como patrones fijos, y apuntan más bien a explotaciones de otros patrones o a usos en decadencia: "El verano volcaba ya oleadas de turismo" (mencionado *supra* como explotación metafórica del patrón 2), "La atención se vuelca en las dos figuras femeninas" (el argumento *atención* no es usual para el patrón 3, "(Alemania) se ha volcado a la fibra de carbono" (no es usual el empleo de la preposición *a* para el patrón 4), etc.

En el caso del primer patrón, por ejemplo, el algoritmo debe identificar, en primer lugar, que el patrón tiene un argumento que se corresponde con el sujeto (es una estructura intransitiva). Identificará como sujetos de *volcar* sustantivos como *autobús, autocar, barca, camión, convoy, embarcación, patera, tren*, etc. En la ontología de Hanks, existe el tipo semántico "Vehicle" como subitpo de "Machine", y se subdivide en "Road Vehicle" y "Water Vehicle". En el caso de *volcar*, tanto puede combinarse con un vehículo terrestre como marino, pues las propiedades semánticas de *volcar* se refieren genéricamente a que un objeto se gira hacia un lado y cae por efecto de la gravedad. Así, puede volcar casi cualquier vehículo, siempre que esté sometido a gravedad y esté apoyado en una superficie, sea agua o suelo.

Una vez identificado el argumento, una ontología suficientemente completa debe proporcionar las relaciones hiperonímicas entre cada sustantivo y el tipo semántico correspondiente: debe asociar *barca, embarcación* y *patera* a "Water Vehicle", y el resto a "Road Vehicle". Siendo ambos tipos semánticos dependientes de "Vehicle", el sistema debe poder unir ambos en este único tipo semántico. No hallando otros sustantivos en posición de sujeto que coincidan con otros tipos semánticos superiores en la jerarquía (es decir, al mismo nivel de "Vehicle").

Pueden encontrarse casos de ambigüedad debida tanto a la sintaxis como a la semántica. La frase "El conductor falleció tras volcar el vehículo" es interpretada por los humanos como intransitiva por conocimiento del mundo, pero un sistema automático podría clasi-

ficarla como transitiva y adjudicarle el patrón 2. Es igualmente difícil clasificar los nombres propios en casos como "Nueva Pepita Aurora volcó, pero no se hundió". Sin embargo, como ya se ha advertido (apartado 4), el objetivo consiste en facilitar y agilizar la labor lexicográfica hasta ahora manual, de modo que se obtenga un porcentaje de aciertos lo suficientemente alto como para justificar el procedimiento.

Dejando de lado, pues, estos casos conflictivos, el algoritmo habría identificado que uno de los patrones con los que se expresa *volcar* es intransitivo y consta de un argumento en posición de sujeto que denota vehículos. Así, este patrón se traslada a la base de datos y se vincula con las concordancias que se ajustan a él. El traslado del patrón a la base de datos lexicográfica es el último paso automático para realizar antes de la redacción propiamente dicha. Puede consultarse el artículo de *volcar* en la versión actual del DAELE (web. cit.).

6. Reflexiones finales y trabajo futuro

En esta exposición se ha propuesto un método de extracción automática de patrones léxicos a través del CPA, con fines lexicográficos. El CPA no tiene por qué ser un sistema válido para cualquier finalidad de desambiguación semántica, pero se considera que resulta útil para la confección de diccionarios, pues ha sido concebido principalmente para ello. La aplicación de esta metodología, como también otras aplicaciones automáticas relacionadas con PLN y CPA, están aún bajo estudio y discusión (Cinková et al., 2010; Ide y Wilks, 2006), por lo que una de las contribuciones del trabajo futuro (si bien secundaria para nuestras necesidades) debe ser ayudar a resolver la cuestión sobre si CPA puede ser una herramienta generalizable para PLN.

Las dificultades de la labor por hacer son evidentes, como también el hecho de que un porcentaje relativamente bajo de éxito equivaldría al mismo porcentaje de tiempo y coste económico ahorrados. Como ya se ha indicado, las herramientas de análisis de corpus disponibles (apartado 2) son muy adecuadas para la tarea lexicográfica, pero está aún por explorar una mayor relación corpus-diccionario.

El plan de implementación de este método se incluye dentro de las actividades para realizar durante el periodo 2011-2012, destinado entre otras cosas a la automatización de procesos para aumentar el rigor y el rendimiento. Se persigue, en definitiva, dotar al diccionario de más herramientas automáticas tanto para el proceso de preparación del contenido como para la interfaz de usuario. Está previsto obtener los primeros resultados en el mismo período.

Bibliografía

Apresjan, J. D. 2002. Principles of systematic lexicography. En M.-H. Corréard, editor, *Lexicograhpy and Natural Language Processing*. Euralex, United Kingdom, páginas 91–104.

Atkins, S. y M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford University Press, United Kingdom.

Battaner, P. y I. Renau. 2008. Sobre las construcciones pronominales y su tratamiento en algunos diccionarios monolingües de cuatro lenguas románicas. En DeCesaris y Bernal (DeCesaris y Bernal, 2008), páginas 495–504.

Boogards, P. 2010. Dictionaries and Second Language Acquisition. En Dykstra y Schoonheim (Dykstra y Schoonheim, 2010), páginas 99–123.

Bosque, I. 2006. Una nota sobre la relevancia de la información sintáctica en el diccionario. En E. Bernal y J. DeCesaris, editores, *Palabra por palabra. Estudios ofrecidos a Paz Battaner*. Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona, páginas 47–53.

Carreras, X., I. Chao, L. Padró, y M. Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers. En *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisboa. European Language Resources Assotiation.

Cinková, S., M. Holub, P. Richlý, L. Smejkalová, y J. indlerová. 2010. Can Corpus Pattern Analysis be used in natural language processing? En *Text, Speech and Dialogue: 13th International Conference*, páginas 67–74, Berlin. Springer.

Cinková, S., M. Holub, y L. Smejkalová. 2010. The lexical population of semantic types in hank's pdev. En G.-M. De

Schryver, editor, *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks*. Menha Publishers, Kampala (Uganda), páginas 199–214.

Convery, C., S. Atkins, A. Kilgarriff, M. Rundell, P. Ó Mianáin, y M. Ó Raghallaigh. 2010. The DANTE Database (Database of Analysed Texts of English). En Dykstra y Schoonheim (Dykstra y Schoonheim, 2010).

Corréard, M.-H. 2002. *Lexicograhpy and Natural Language Processing*. Euralex, United Kingdom.

DeCesaris, J. y E. Bernal, editores. 2008. *Proceedings of the XIII EURALEX International Congress*, Barcelona. Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Dykstra, A. y T. Schoonheim, editores. 2010. *Proceedings of the XIV Euralex International Congress*, Ljowvert (Países Bajos). Fryske Akademy.

Fontenelle, T. 2002. Lexical knowledge and natural language processing. En M.-H. Corréard, editor, *Lexicograhpy and Natural Language Processing*. Euralex, United Kingdom, páginas 216–229.

Hanks, P. 2004. The Syntagmatics of Metaphor and Idioms. *International Journal of Lexicography*, 17(3):245–274.

Hanks, P. En prensa. *Lexical Analysis: Norms and Exploitations*. MIT Press, Massachusetts.

Hanks, P. y J. Pustejovsky. 2000. A Pattern Dictionary for Natural Language Processing. *Revue Française de Lingüistique Apliquée*, 10(2):63–82.

Hjelmslev, L. 1943. *Prolegomena to a Theory of Language*. Indiana University Publications in Anthropology and Linguistics, Baltimore.

Ide, N. y Y. Wilks. 2006. Making senses about sense. En E. Agirre y P. Edmonds, editores, *Word Sense Disambiguation. Algorithms and Applications*. Springer, Nueva York, páginas 47–74.

J. Sinclair, dir. 1987. *The Collins Cobuild English Language Dictionary*. Harper-Collins, Nueva York.

Jezek, E. y F. Frontini. 2010. From pattern dictionary to patternbank. En G.-M. De Schryver, editor, *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks*. Menha Publishers, Kampala (Uganda), páginas 215–239.

Kilgarriff, A., P. Rhychlý, P. Smrz, y P. Tugwell. 2004. The Sketch Engine. En *Proceedings of the Eleventh Euralex International Congress*, páginas 105–116, Lorient (Francia). Université de Bretagne-Sud.

Miller, G. A., Ch. Fellbaum, J. Kegl, y J. K. Miller. 1988. WordNet: an electronic lexical reference system based on theories of lexical memory. *Revue Quebecoise de Linguistique*, 17(2):181–213.

Nazar, R. y M. Janssen. 2010. Combining resources: taxonomy extraction from multiple dictionaries. En *Proceedings of LREC 2010 (The 8th edition of the Language Resources and Evaluation Conference)*, páginas 1055–1061, Valletta (Malta).

Nazar, R., J. Vivaldi, y L. Wanner. En prensa. Automatic taxonomy extraction for specialized domains using distributional semantics. *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication*.

Pustejovsky, J. 1995. *The Generative Lexicon*. MIT Press, Massachusetts.

RAE. 2009. *Nueva gramática de la lengua española*. Espasa, Madrid.

Renau, I. y A. Alonso. En preparación. Using Corpus Pattern Analysis for the Spanish Learner's Dictionary DAELE (Diccionario de aprendizaje del español como lengua extranjera). En *Corpus Linguistics 2011*, Birmingham.

Sinclair, J. 1998. The lexical item. En E. Wiegand, editor, *Contrastive Lexical Semantics*. John Benjamins, Amsterdam, páginas 1–14.

Wilks, Y., B. Slator, y L. Guthrie. 1996. *Electric Words. Dictionaries, Computers and Meanings*. MIT Press-Cambridge, Massachusetts-Cambridge.