

XXIII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural



Universidad de Sevilla
10, 11 y 12 de septiembre de 2007



indisys:

Intelligent Dialogue Systems

Departamento de
Lenguajes y Sistemas Informáticos
Universidad de Sevilla



escuela técnica superior de ingeniería informática
Universidad de Sevilla

EDITORES

Víctor J. Díaz Madrigal (Univ. de Sevilla)
Fernando Enríquez de Salamanca Ros (Univ. de Sevilla)

COMITÉ CIENTÍFICO

PRESIDENTE

Prof. Víctor Jesús Díaz Madrigal (Universidad de Sevilla)

MIEMBROS

Prof. José Gabriel Amores Carredano (Universidad de Sevilla)
Prof. Toni Badia i Cardús (Universitat Pompeu Fabra)
Prof.^a Irene Castellón Masalles (Universitat de Barcelona)
Prof. Manuel de Buenaga Rodríguez (Universidad Europea de Madrid)
Prof. Ricardo de Córdoba (Universidad Politécnica de Madrid)
Prof.^a Arantza Díaz de Ilaraza (Euskal Herriko Unibertsitatea)
Prof. Antonio Ferrández Rodríguez (Universitat d'Alacant)
Prof. Mikel Forcada Zubizarreta (Universitat d'Alacant)
Prof.^a Ana María García Serrano (Universidad Politécnica de Madrid)
Prof. Koldo Gojenola Gallettebeitia (Euskal Herriko Unibertsitatea)
Prof. Xavier Gómez Guinovart (Universidade de Vigo)
Prof. Julio Gonzalo Arroyo (Universidad Nacional de Educación a Distancia)
Prof. José Miguel Goñi Menoyo (Universidad Politécnica de Madrid)
Prof. Ramón López-Cózar Delgado (Universidad de Granada)
Prof. Javier Macías Guarasa (Universidad Politécnica de Madrid)
Prof. José B. Mariño Acebal (Universitat Politècnica de Catalunya)
Prof.^a M. Antonia Martí Antonín (Universitat de Barcelona)
Prof.^a Raquel Martínez (Universidad Nacional de Educación a Distancia)
Prof. Antonio Molina Marco (Universitat Politècnica de Valencia)
Prof. Juan Manuel Montero (Universidad Politécnica de Madrid)
Prof.^a Lidia Ana Moreno Boronat (Universitat Politècnica de Valencia)
Prof. Lluís Padró (Universitat Politècnica de Catalunya)
Prof. Manuel Palomar Sanz (Universitat d'Alacant)
Prof. Germán Rigau (Euskal Herriko Unibertsitatea)
Prof. Horacio Rodríguez Hontoria (Universitat Politècnica de Catalunya)
Prof. Emilio Sanchís (Universitat Politècnica de Valencia)
Prof. Kepa Sarasola Gabiola (Euskal Herriko Unibertsitatea)
Prof. L. Alfonso Ureña López (Universidad de Jaén)
Prof. Ferrán Pla (Universitat Politècnica de Valencia)
Prof.^a M^a Felisa Verdejo Maillo (Universidad Nacional de Educación a Distancia)
Prof. Manuel Vilares Ferro (Universidade de Vigo)

Revisores Externos

Iñaki Alegria, Laura Alonso Alemany, Kepa Bengoetxea, Zoraida Callejas Carrión, Francisco Carrero, Vicente Carrillo Montero, Fermín Cruz Mata, Víctor Manuel Darriba Bilbao, César de Pablo Sánchez, Fernando Enríquez de Salamanca Ros, Milagros Fernández Gavilanes, Ana Fernández Montraveta, Óscar Ferrández, Sergio Ferrández, Miguel Ángel García Cumbreñas, Manuel García Vega, Rubén Izquierdo Beviá, Zornitsa Kozareva, Sara Lana Serrano, Mikel Lersundi, Lluís Márquez, María Teresa Martín Valdivia, José Luis Martínez Fernández, Germán Montoro Manrique, Andrés Montoyo Guijarro, Iulia Nica, Francisco Javier Ortega Rodríguez, Jesús Peral Cortés, Enrique Puertas, Francisco José Ribadas Pena, Estela Saquete Boró, José Antonio Troyano Jiménez, Gloria Vázquez.

COMITÉ ORGANIZADOR

PRESIDENTE

Víctor Jesús Díaz Madrigal

MIEMBROS

Adolfo Aumaitre del Rey

Rafael Borrego Roper

José Miguel Cañete Valdeón

Vicente Carrillo Montero

Fermín Cruz Mata

Fernando Enríquez de Salamanca Ros

Francisco José Galán Morillo

Carlos García Vallejo

Fco. Javier Ortega Rodríguez

Luisa María Romero Moreno

José Antonio Troyano Jiménez

Preámbulo

El ejemplar número 39 de la revista de la Sociedad Española para el Procesamiento del Lenguaje Natural contiene los artículos científicos - más los resúmenes de proyectos de investigación y de demostraciones de herramientas - aceptados por el Comité Científico para su presentación en el XXIII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN'07). Esta edición del congreso ha sido organizada por miembros del departamento de Lenguajes y Sistemas Informáticos de la Universidad de Sevilla en la Escuela Técnica Superior de Ingeniería Informática. El número de artículos de investigación recibido junto con la continuidad en la celebración anual del congreso, ésta es la vigésimo tercera edición ininterrumpida, no hacen más que constatar el interés y la actualidad que disfruta hoy en día la investigación en el campo de las Tecnologías de la Lengua.

Estas actas recogen 32 artículos científicos que podemos agrupar de forma no categórica y excluyente en las siguientes áreas temáticas: Análisis Morfosintáctico (4 trabajos), Búsqueda de Respuestas (2 trabajos), Categorización de Textos (3 trabajos), Extracción de Información (5 trabajos), Lexicografía Computacional (4 trabajos), Lingüística de Corpus (4 trabajos), Semántica (4 trabajos), Sistemas de Diálogo (2 trabajos) y Traducción Automática (4 trabajos). Se recibieron un total de 49 trabajos de los cuales tan sólo las 32 contribuciones mencionadas (65 por ciento) obtuvieron la aprobación global del Comité Científico. Cada uno de los trabajos recibidos fue revisado por 3 miembros del Comité Científico. Además, y como viene siendo habitual, en las actas se incluyen dos resúmenes presentando proyectos de investigación y nueve resúmenes presentando demostraciones de herramientas de uso específico para tareas relacionadas con el Procesamiento del Lenguaje Natural.

Esta edición del congreso cuenta con 2 conferencias invitadas a cargo del Dr. D. Antal van den Bosch (Universidad de Tilburg) y del Dr. D. Anselmo Peñas (Universidad Nacional de Educación a Distancia). Este año se da la peculiaridad de que durante los días 11 y 12 de septiembre, en paralelo con el congreso, se celebran las Jornadas de la Red Temática para el Tratamiento de la Información Multilingüe y Multimodal. En el seno de dichas jornadas se incluye la conferencia invitada a cargo del Dr. D. Ralf Steinberger (Joint Research Centre).

No quiero acabar estas líneas sin dar las gracias a los patrocinadores del congreso ya que sin su apoyo financiero o logístico hubiera sido muy difícil organizarlo. No puedo tampoco dejar de agradecer el esfuerzo y las facilidades de las que he sido objeto por parte de todos los miembros del Comité Científico y del Órgano de Gobierno de la Sociedad. Finalmente, me gustaría acabar recordando a todos mis compañeros del grupo de investigación ITALICA por el trabajo adicional que ha supuesto la preparación de este evento.

Víctor Jesús Díaz Madrigal
Presidente del Comité de Programa de XXIII Congreso de la SEPLN



Sociedad Española para el
Procesamiento del Lenguaje Natural

ARTÍCULOS

Análisis Morfosintáctico

<i>Desarrollo de un Analizador Sintáctico Estadístico basado en Dependencias para el Euskera</i> Kepa Bengoetxea y Koldo Gojenola	5
<i>Técnicas Deductivas para el Análisis Sintáctico con Corrección de Errores</i> Carlos Gómez-Rodríguez, Miguel A. Alonso y Manuel Vilares	13
<i>A Simple Formalism for Capturing Order and Co-Occurrence in Computational Morphology</i> Mans Hulden y Shannon Bischoff	21
<i>A Note on the Complexity of the Recognition Problem for the Minimalist Grammars with Unbounded Scrambling and Barriers</i> Alexander Perekrestenko	27

Búsqueda de Respuestas

<i>Paraphrase Extraction from Validated Question Answering Corpora in Spanish</i> Jesús Herrera, Anselmo Peñas y Felisa Verdejo	37
<i>Evaluación de Sistemas de Búsqueda de Respuestas con restricción de tiempo</i> Fernando Llopis, Elisa Noguera, Antonio Ferrández y Alberto Escapa	45

Categorización de Textos

<i>Medidas Internas y Externas en el Agrupamiento de Resúmenes Científicos de Dominios Reducidos</i> Diego Ingaramo, Marcelo Errecalde y Paolo Rosso	55
<i>Integración de Conocimiento en un Dominio Específico para Categorización Multietiqueta</i> María Teresa Martín, Manuel Carlos Díaz, Arturo Montejo y L. Alfonso Ureña-López	63
<i>Similitud entre Documentos Multilingües de Carácter Científico-Técnico en un Entorno Web</i> Xabier Saralegi y Iñaki Alegria	71

Extracción de Información

<i>The Influence of Context during the Categorization and Discrimination of Spanish and Portuguese Person Names.</i> Zornitsa Kozareva, Sonia Vázquez y Andrés Montoyo	81
<i>Studying CSSR Algorithm Applicability on NLP Tasks</i> Muntsa Padró y Lluís Padró	89
<i>Aprendizaje Automático para el Reconocimiento Temporal Multilingüe basado en TiMBL</i> Marcel Puchol-Blasco, Estela Saquete y Patricio Martínez-Barco	97
<i>Alias Assignment in Information Extraction</i> Emili Sapena, Lluís Padró y Jordi Turmo	105
<i>Evaluación de un Sistema de Reconocimiento y Normalización de Expresiones Temporales en Español</i> María Teresa Vicente-Díez, César de Pablo-Sánchez y Paloma Martínez	113

Lexicografía Computacional

<i>Inducción de Clases de Comportamiento Verbal a partir del Corpus SENSEM</i> Laura Alonso, Irene Castellón y Nevena Tinkova	123
<i>An Open-Source Lexicon for Spanish</i> Montserrat Marimon, Natalia Seghezzi y Núria Bel	131
<i>Towards Quantitative Concept Analysis</i> Rogelio Nazar, Jorge Vivaldi y Leo Wanner	139
<i>Evaluación Automática de un Sistema Híbrido de Predicción de Palabras y Expansiones</i> Sira Elena Palazuelos, José Luis Martín y Javier Macías	147

Lingüística de Corpus

<i>Specification of a General Linguistic Annotation Framework and its Use in a Real Context</i>	
Xabier Artola, Arantza Díaz de Ilarraza, Aitor Sologaitoa y Aitor Soroa	157
<i>Determinación del Umbral de Representatividad de un Corpus mediante el Algoritmo N-Cor</i>	
Gloria Corpas y Miriam Seghiri	165
<i>Generación Semiautomática de Recursos</i>	
Fernando Enríquez, José Antonio Troyano, Fermín Cruz y F. Javier Ortega	173
<i>Building Corpora for the Development of a Dependency Parser for Spanish Using Maltparser</i>	
Jesús Herrera, Pablo Gervás, Pedro J. Moriano, Alfonso Muñoz y Luis Romero	181

Semántica

<i>A Proposal of Automatic Selection of Coarse-grained Semantic Classes for WSD</i>	
Rubén Izquierdo-Bevia, Armando Suárez y Germán Rigau	189
<i>Cognitive Modules of an NLP Knowledge Base for Language Understanding</i>	
Carlos Perrián-Pascual y Francisco Arcas-Túnez	197
<i>Text as Scene: Discourse Deixis and Bridging Relations</i>	
Marta Recasens, Antonia Martí Antonín y Mariona Taulé	205
<i>Definición de una Metodología para la Construcción de Sistemas de Organización del Conocimiento a partir de un Corpus Documental en Lenguaje Natural</i>	
Sonia Sánchez-Cuadrado, Jorge Morato, José Antonio Moreiro y Monica Marrero	213

Sistemas de Diálogo

<i>Prediction of Dialogue Acts on the Basis of the Previous Act</i>	
Sergio R. Coria y Luis Alberto Pineda	223
<i>Adaptación de un Gestor de Diálogo Estadístico a una Nueva Tarea</i>	
David Griol, Lluís F. Hurtado, Encarna Segarra y Emilio Sanchís	231

Traducción Automática

<i>Un Método de Extracción de Equivalentes de Traducción a partir de un Corpus Comparable Castellano-Gallego</i>	
Pablo Gamallo y José Ramon Pichel	241
<i>Flexible Statistical Construction of Bilingual Dictionaries</i>	
Ismael Pascual y Michael O'Donnell	249
<i>Training Part-of-Speech Taggers to build Machine Translation Systems for Less-Resourced Language Pairs</i>	
Felipe Sánchez-Martínez, Carme Armentano-Oller, Juan Antonio Pérez-Ortiz y Mikel L. Forcada	257
<i>Parallel Corpora based Translation Resources Extraction</i>	
Alberto Simões y José João Almeida	265

DEMOSTRACIONES

<i>Una Herramienta para la Manipulación de Corpora Bilingüe usando Distancia Lexica</i>	
Rafael Borrego y Víctor J. Díaz	275
<i>MyVoice goes Spanish. Cross-lingual Adaptation of a Voice Controlled PC Tool for Handicapped People</i>	
Zoraida Callejas, Jan Nouza, Petr Cerva y Ramón López-Cózar	277
<i>HistoCat y DialCat: Extensiones de un Analizador Morfológico para tratar Textos Históricos y Dialectales del Catalán</i>	
Jordi Duran, M ^a Antonia Martí y Pilar Perea	279
<i>MorphOz: Una Plataforma de Desarrollo de Analizadores Sintáctico-Semánticos Multilingüe</i>	
Oscar García	281
<i>Sistema de Diálogo Estadístico y Adquisición de un Nuevo Corpus de Diálogos</i>	
David Griol, Encarna Segarra, Lluís F. Hurtado, Francisco Torres, María José Castro, Fernando García y Emilio Sanchís	283
<i>JBeaver: Un Analizador de Dependencias para el Español</i>	
Jesús Herrera, Pablo Gervás, Pedro J. Moriano, Alfonso Muñoz y Luis Romero	285
<i>NowOnWeb: a NewsIR System</i>	
Javier Parapar y Álvaro Barreiro	287
<i>The Coruña Corpus Tool</i>	
Javier Parapar y Isabel Moskowich-Spiegel	289
<i>WebJspell, an Online Morphological Analyser and Spell Checker</i>	
Rui Vilela	291

PROYECTOS

<i>El Proyecto Gari-Coter en el Seno del Proyecto RICOTERM2</i>	
Fco. Mario Barcala, Eva Domínguez, Pablo Gamallo, Marisol López, Eduardo Miguel Moscoso, Guillermo Rojo, María Paula Santalla del Río y Susana Sotelo	295
<i>Portal da Lingua Portuguesa</i>	
Maarten Janssen	297

El proyecto Gari-Coter* en el seno del proyecto RICOTERM2**

Fco. Mario Barcala Rodríguez y Eva M.^a Domínguez Noya
 Centro Ramón Piñeiro para a Investigación en Humanidades
 {fbarcala,edomin}@cirp.es

Pablo Gamallo Otero y Marisol López Martínez y Eduardo Miguel Moscoso Mato y
 Guillermo Rojo y María Paula Santalla del Río y Susana Sotelo Docío
 Universidade de Santiago de Compostela
 {pablogam, fgmarsol, fgmato, guillermo.rojo, fescocio}@usc.es

Resumen: Descripción del proyecto Gari-Coter para la elaboración de los recursos lingüísticos en gallego necesarios para un re-elaborador de consultas multilingüe.

Palabras clave: expansión de consultas, corpus, base de datos terminológica, extracción automática de términos

Abstract: Description of the Gari-Coter project for the development of the necessary linguistic resources in Galician for a multilingual query re-elaborator.

Keywords: query expansion, corpus, terminological database, automatic terminology extraction

1. Situación actual

Como se ha indicado en la nota de agradecimiento adjunta al acrónimo del proyecto incluido en el título, éste se ha venido desarrollando desde 2004, y su cierre está previsto para finales de 2007. Dos años y medio, por tanto, lleva el proyecto en curso, por lo cual lo que incluimos aquí es una presentación esquemática de lo que se proponía, así como de algunos de sus, ahora ya, resultados de hecho, a falta de un sexto de tiempo de desarrollo del proyecto. Lo que queda del mismo, por otra parte, es previsible que se dedique a la integración de los recursos y herramientas generados en el seno de cada uno de los subproyectos que integran el proyecto coordinado RICOTERM2, el propio Gari-Coter, y el subproyecto, del mismo nombre que el coordinado, RICOTERM2¹.

2. El subproyecto Gari-Coter en el seno del proyecto coordinado RICOTERM2

El proyecto coordinado RICOTERM2 tiene como objetivo principal el desarrollo de un prototipo para un sistema multilingüe de reformulación de consultas planteadas por usuarios de Internet interesados en la búsqueda de información acerca de un ámbito comunicativo especializado, en nuestro caso, economía. El sistema se integrará, como se describe en (Lorente, 2005), en una aplicación que consistirá en una interfaz, ubicada en un portal web especializado en economía, para la transformación de consultas simples en consultas multilingües expandidas lingüística y conceptualmente. Actualmente las lenguas de trabajo son el catalán, el castellano, el gallego, el inglés y el vasco. El diseño general del prototipo está también descrito en (Lorente, 2005): baste aquí, para que puedan ser cabalmente entendidos los objetivos específicos del subproyecto Gari-Coter, indicar que, con el propósito de mejorar los resultados de las aplicaciones implicadas de Recuperación de Información mediante técnicas de expansión de consultas, el proyecto utiliza métodos tanto de expansión únicamente por términos (*only-term expansion*) como de expansión de texto completo (*full-text expansion*). Para lo primero, se hará uso de una ontología del dominio. Para lo segundo, de un corpus específico de economía, estructural y lingüísticamente

* *Creación e integración multilingüe de recursos terminológicos en gallego para Recuperación de Información mediante estrategias de control terminológico y discursivo en ámbitos comunicativos especializados.* Subproyecto financiado, bajo la dirección de M.^a Paula Santalla, por el Ministerio de Educación y Ciencia entre 2004 y 2007 (HUM2004-05658-C02-02/FILO).

** *Control terminológico y discursivo para la recuperación de información en ámbitos comunicativos especializados, mediante recursos lingüísticos específicos y un reelaborador de consultas.* Proyecto coordinado financiado, bajo la dirección de Mercè Lorente Casafont, por el Ministerio de Educación y Ciencia entre 2004 y 2007 (HUM2004-05658-C02-00/FILO).

te anotado, el cual habrá de servir para, mediante el recurso a herramientas como extractores automáticos de terminología y similares, detectar colocaciones o fraseología propia de los términos introducidos por el propio usuario, u obtenidos tras la consulta a la ontología.

Dentro de este planteamiento general, el proyecto Gari-Coter (aparte de objetivos compartidos, relacionados, como puede suponerse, con el diseño y la integración de todo lo producido en una aplicación web) tiene como objetivos propios la constitución de los recursos para el gallego: un corpus de economía, adecuadamente codificado y anotado, adaptando para ello herramientas de procesamiento existentes para el gallego, y un banco de datos terminológicos, obtenido a partir de recursos previos y de la explotación del propio corpus constituido. A falta de algo más de seis meses para la finalización del proyecto, estos recursos han podido ser elaborados en la forma y dimensión que someramente describimos a continuación.

2.1. El corpus

Como para todas las lenguas implicadas en el proyecto RICOTERM2, no uno sino, en realidad, dos subcorpus de dominio han sido desarrollados para el gallego: un subcorpus genérico y uno específico. El primero integrado por 609 noticias de periódico que suman 206510 palabras distribuidas en 7892 oraciones. El segundo integrado por 14 libros y dos revistas especializadas que entre todos suman 801702 palabras distribuidas en 34588 oraciones.

Ambos corpus están codificados utilizando el estándar XML. Cada documento consta de una cabecera con información bibliográfica y de contenido, seguida ésta del documento mismo, estructurado hasta el nivel de la oración. Ambos corpus, asimismo, han sido anotados morfosintácticamente con información acerca de clase de palabras y categorías flexivas consideradas relevantes.

En línea con los planteamientos generales del proyecto coordinado (búsqueda y aprovechamiento de recursos preexistentes), para la constitución de ambos corpus llegamos a un acuerdo con el Centro Ramón Piñeiro para a Investigación en Humanidades², que nos cedió los textos procedentes del corpus CORGA, Corpus de Referencia del Gallego Actual, procesados lingüísticamente con su pro-

pio sistema de etiquetación. Toda la anotación del corpus genérico fue corregida manualmente.

2.2. El banco de datos terminológico

El banco de datos terminológico se ha elaborado a partir, por un lado, de recursos previos que constituían fuentes considerablemente heterogéneas³ en cuanto a calidad, dimensión y fiabilidad: dos diccionarios, dos glosarios electrónicos y la sección de economía de una base de datos terminológica, ésta última la más rica y rigurosa sin duda.

Actualmente, el banco de datos consta de 6046 términos del dominio económico obtenidos por esta vía, la mayoría de ellos asociados a información exhaustiva acerca del lema, la clase de palabras y la definición, así como, en la mayoría de los casos, equivalentes en otras lenguas e información sobre sinónimos e hiperónimos.

El conjunto de términos descrito, así como el corpus, se han utilizado además para, mediante técnicas de extracción automática de términos multipalabra basadas en medidas de similitud contextual, ampliar el banco de datos terminológico. En la última de las experiencias llevadas a cabo 740 términos multipalabra pudieron obtenerse, pero los resultados de precisión asociados, debidos sin duda al reducido tamaño del corpus, aconsejan, cuanto menos, una revisión manual de los mismos.

Notas

¹Con el mismo acrónimo y nombre que el proyecto coordinado, financiado por el Ministerio de Educación y Ciencia entre 2004 y 2007, y dirigido por Mercè Lorente (HUM2004-05658-C02-01/FILO).

²<http://www.cirp.es>. [Consultado: 6, junio, 2007].

³**Eiras:** Eiras Rey, A.: *Diccionario de economía*, no publicado. **Formoso:** Formoso Gosende, V. (coord.) (1997): *Diccionario de términos económicos e empresariales galego-castelán-inglés*. Santiago de Compostela: Confederación de Empresarios de Galicia. **Panlatin Electronic Commerce Glossary:** <http://fon.gs/panlatino>. **Glossary about commerce from galego.org:** <http://galego.org/vocabularios/ccomercial.html>. **SNL:** <http://www.usc.es/en/servizos/portadas/snl.jsp>.

Bibliografía

- Lorente, M. 2005. Ontología sobre economía y recuperación de información [en línea]. *Hipertext.net*, (3). <http://www.hipertext.net>. [Consultado: 30, enero, 2007].