

Text-to-speech --- a rewriting system approach 247
 José Jodó Almeida, Alberto Manuel Simóns

Una nueva técnica para evaluar sistemas conversacionales basada en la generación automática de diálogos 255
 R. López-Cózar, J. C. Segura, A. De la Torre, A. J. Rubio

Categorización de textos multilingües basada en Redes Neuronales 265
 Manuel García Vega, Marie Martin Valdivia, L. Alfonso Ureña López

Cross-lingual keyword assignment 273
 Rafi Sternberger

Generación automática de resúmenes personalizados 281
 Ignacio Acero, Matías Alcalá, Alberto Díaz, José María Gómez, Manuel Maño

Proyectos 291
 Proyecto europeo D'Homme
 José F. Quesada, J. Gabriel Amores

XML-Bi: Procesamientos para gestión de flujo documental multilingüe sobre XML/TEI 293
 Josefa Abaitua, Arantza Dominguez, Carmen Isasi, José Luis Ramirez, Inés Jacob, Idaira Madariaga, Arantza Castiella, Raquel Martínez, Alberto Garay, Thomas Diedrich

Proyecto de indexado automático para documentos en el campo de la física de altas energías 295
 Arturo Montijo Rábiz

Proyecto Tag parsing 297
 J. Gabriel Amores

Hermes: Servicios de personalización inteligente de noticias mediante la integración de técnicas de análisis automático del contenido textual y modelado de usuario con capacidades bilingües 299
 Alberto Díaz, Manuel de Buenaño, Inés y Giráldez, José María Gómez, Antonio García, Inmaculada Chacón, Beatriz San Miguel, Enrique Puertas, Raúl Mauriano, Matías Alcalá, Ignacio Acero, Pablo Gervás

Spanish Acquisition: Analysis of learner corpora generated through inter-cultural telecollaboration 301
 Julia Kasher, James Lantolf, Steven Thorne, Antonio Jiménez, Brenda Ross, Sagrario Salaberri

Proyecto europeo Sirdus 303
 José F. Quesada, J. Gabriel Amores

XTRA-Bi: Extracción automática de entidades bitextuales para software de traducción asistida 305
 Inés Jacob, Joseba Abaitua, Jonaki Diaz, José Gómez, Koldo Otxina

Demostraciones 309
 Análisis y expansión de consultas en lenguaje natural para mejora de la búsqueda en Web
 Alberto Ruiz, Paloma Martínez, Ana García-Serrano

ANTRO: Un sistema de reconocimiento y gestión de antropónimos 311
 Daniel Casanova, Xavier Lloré, Rafael Marín, Josep M. Mercenario, Genar Pérez, David Tratzig

Diccionario electrónico de sinónimos y antónimos de la lengua española (DESALLE) 313
 Santiago Fernández Lanza

EGEO: La estructura geográfica de una base de conocimiento 315
 Ignacio González, Sergi Cornell, Josep M. Mercenario, David Tratzig, Juan Vaqué

El uso de editoriales de herramientas lingüísticas: un ejemplo con los descriptores humanos 317
 Adán Casan, Sergi Cornell, Miriam Colom, Mireia Farrus, Ignacio González, Rafael Marín, David Tratzig

Gestión flexible de diálogos en el proyecto ADVICE 319
 Luis Rodrigo Aguado, Ana García Serrano, Paloma Martínez

Lajuj: Un sistema de recuperación multilingüe basado en EuroWordNet 321
 Fernando Martínez Santiago, Manuel Carlos Díaz Galisano, L. Alfonso Ureña López, Maic Martín Valdivia, Manuel García Vega, José Ramón Balsas Almagro

Procesamiento del Lenguaje Natural, Revista nº 27, septiembre 2001

Artículos 13
 Aspectos ortográficos, léxicos y morfológicos del etiquetado lingüístico de un corpus de informática en lengua gallega
 José Luis Aguirre Moreno, Nuria Andión Rodríguez, Xavier Gómez Guillot

Creación, etiquetación y desambiguación de un corpus de referencia del español 21
 Montserrat Llad Ferrasella, Irene Castellón Masalles, M. Antonia Martí Antonia

Extracción masiva de información sobre subcategorización verbal vasca a partir de corpus 29
 J. Aldazabal, M. Aranzubi, A. Atutza, K. Goyenola, K. Sarasola, P. Arriaga

Problemática de la recogida y anotación de una base de datos oral para el gallego 37
 Begonia González Rev, Antonio Cerdenal López, Laura Ducio Fernández, Carmen García Mateo

Análisis sintáctico ascendente de TAGs guiado por la esquina izquierda 47
 Vicente Carrillo Montero, Víctor J. Díaz Madruga, Miguel A. Alonso Pardo

Una aproximación para resolución de ambigüedad estructural empleando tres mecanismos diferentes 55
 Sofía N. Galicia-Haro, Alexander F. Gelbukh, Igor A. Bolshakov

Corpus-based stochastic finite-state predictive text entry for reduced keyboards application to Catalan 65
 Mikel L. Forcadell

Integration of dialogue moves and speech recognition in a telephone scenario 71
 José F. Quesada, J. Gabriel Amores, Rafael Ballarín

Dialogue moves for natural command languages 81
 J. Gabriel Amores, José F. Quesada

Dialogue management in a home machine environment: linguistic components over an agent architecture 89
 José F. Quesada, Federico García, Esther Sent, José Angel Bernal, Gabriel Amores

Propuesta de un espacio de accesibilidad anafórica estructural para textos HTML 97
 Borja Navarro, Patricia Martínez-Buena, Rafael Mañón

Definición de un modelo semántico aplicado a los sistemas de búsqueda de respuestas 107
 José Luis Toledo González, Antonio Fernández Rodríguez

Un método de agrupamiento de grafos conyugales para minería de texto 115
 M. Montes-Gómez, A. Gelbukh, A. López-López, R. Biezas-Liós

Normalización de términos multilingües mediante pares de dependencia sintáctica 123
 Jesús Flores, Feo. Mario Barco, Miguel A. Alonso

Un modelo de recuperación de información basado en redes bayesianas 131
 Luis M. de Campos, Juan F. Huete, Juan M. Fernández-Lara

WWW como fuente de recursos lingüísticos para su uso en PLN 141
 Fernando Martínez Santiago, L. Alfonso Ureña López, Manuel García Vega

El sistema de traducción automática castellano <-> catalán INTERSTRUM Alicante 151
 R. Conal-Morán, A. Escriba-Guillem, A. Garrido-Alcázar, M.J. González-Sorvil, Ferrnand-Bellver, S. Montserrat-Buena, S. Oute-Rojas, H. Pastor-Pina, P.M. Poye-Antón, M.L. Forcadell

MorphTrans: un lenguaje y un compilador para especificar y generar módulos de transferencia morfológica para sistemas de traducción automática 157
 Alicia Garrido-Alcázar, Mikel L. Forcadell

Extracción de relaciones léxico-semánticas a partir de palabras derivadas usando patrones de definición 165
 Enckó Agorri, Mikel Leraunak

Etiquetación robusta del lenguaje natural: preprocesamiento y segmentación 173
 Jorge Grau Gil, Feo. Mario Barco Rodríguez, Jesús Vilares Pardo

Generación automática de familias morfológicas mediante morfología derivativa productiva 181
 Jesús Vilares, David Cuervo, Miguel A. Alonso

Internet como fuente de información léxica: extracción de etiquetas de dominio y detección de nuevos sentidos 189
 Céfira Santamaría, Julio González Arroyo

A POS-Tagger generator for unknown languages 199
 Nuno C. Marques, Gabriel Pereira Lopes

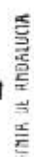
Estudio de cooperación de métodos de desambiguación léxica: marcas de especificidad vs. máxima entropía 207
 Armando Suárez, Andrés Mantoyo

Evaluación de un etiquetador morfológico basado en bigramas especializados para el castellano 215
 Ferrn Pila, Antonio Molina, Nuria Andión

Asignación automática de marcas de pitch basada en programación dinámica 225
 Francesc Alías Pujol, Ignasi Frando Sanz

Modelo cuantitativo de enonación del español 233
 David Escudero Mancebo, Valentín Cardenoso Payo

Transcriptor ortográfico-fonético para el castellano 241
 María José Castro, Salvador España, Ismael Salvador, Andrés Marsala



EDITADO POR:

L. Alfonso Ureña López (Universidad de Jaén)

COMITÉ DE PROGRAMA:

Presidente:

L. Alfonso Ureña López

Miembros:

1. Modelos lingüísticos, matemáticos y psicolingüísticos del lenguaje
Horacio Rodríguez (Universidad Politécnica de Cataluña)
A. Martí (Universidad de Barcelona)
2. Lingüística de corpus
Xavier Gómez (Universidad de Vigo)
Joseba Abaitua (Universidad de Deusto)
3. Extracción y recuperación de información
Julio Gonzalo (U.N.E.D)
José M. Goñi (Universidad Politécnica de Madrid)
4. Gramáticas y formalismos para el análisis morfológico y sintáctico
Manuel Vilares (Universidad de La Coruña)
Koldo Gojenola (Universidad de País Vasco)
5. Lexicografía computacional
Irene Castellón (Universidad de Barcelona)
Toni Badia (Universidad Pompeu Fabra)
6. Generación textual monolingüe y multilingüe
Gabriel Amores (Universidad de Sevilla)
7. Traducción automática
Kepa Sarasola (Universidad del País Vasco)
Antonio Fernández (Universidad de Alicante)
8. Reconocimiento y síntesis de voz
Nati Prieto (Universidad Politécnica de Valencia)
9. Semántica, pragmática y discurso
Ana García-Serrano (Universidad Politécnica de Madrid)
10. Resolución de la ambigüedad léxica
L. Alfonso Ureña (Universidad de Jaén)
Manuel Palomar (Universidad de Alicante)
11. Recuperación de información multilingüe
Felisa Verdejo (U.N.E.D)
Lidia Moreno (Universidad de La Coruña)
12. Aplicaciones industriales del PLN
Lluís Padró (Universidad Politécnica de Cataluña)
13. Análisis automático del contenido textual
Manuel de Buena (Universidad Europea de Madrid)

COMITÉ DE ORGANIZACIÓN:

Presidente:
L. Alfonso Ureña López (Universidad de Jaén)

Secretario:
Manuel García Vega (Universidad de Jaén)

Miembros:
M^a Teresa Martín Valdivia (Universidad de Jaén)
Fernando Martínez Santiago (Universidad de Jaén)
Manuel Carlos Díaz Gallano (Universidad de Jaén)
José Ramón Balsas (Universidad de Jaén)
Pedro González García (Universidad de Jaén)
Victor Rivas Santos (Universidad de Jaén)

REVISORES EXTERNOS:

Eneko Agirre Bengoa
Guadalupe Aguado
Pablo Albar Ausina
Iraki Alegria Loinaz
Manuel Alonso González
Miguel A. Alonso Pardo
Margarita Alonso Ramos
Alberto Alvarez Lugiis
Montserrat Arévalo Rodríguez
María Victoria Arranz Corzana
David Cabriero Souto
José Carlos González
Juan Carlos Pérez
Xavier Carreiras Pérez
Núria Castells Ariño
María José Castro Bleda
Montserrat Civit Torruella
Salvador Climent Roca
Arantza Díaz de Ibarraza
Victor J. Díaz Madrigal
Nerea Ezeiza Ramos
Gregorio Fernández Fernández
Carlos Figuerola
Mikel Forcada
Pablo de la Fuente Redondo
Manuel García Vega
Ignacio Giráldez
Jorge Graña Gil

Ángels Hernández Gómez
Eduardo Lleida Solano
Joaquín Llistern
Fernando Llopis
Luis Márquez Vilodre
M. Antonia Martí Antonín
Paloma Martínez Fernández
Andrés Montoyo Gujarró
Antonio Moreno Sandoval
Javier Pérez Guerra
Ferran Pla Santamaria
Celia Rico Pérez
German Rigau
Fernando Saenz Pérez
Fernando Sánchez León
Antonio Sánchez Valderribanos
Emilio Sanchis Arnal
Encarna Segarra Soriano
María José Simón Aragón
Alejandro Sobrino Cerdobitúa
Armando Suárez Cuelo
Declerck Thierry
Antonio Valderribanos
Gloria Vázquez
Julio Villena Román
Jorge Vivaldi

Artículos

Lingüística del corpus 11

Aspectos ortográficos, léxicos y morfosintácticos del etiquetado lingüístico de un corpus de informaticos en lengua gallega
José Luis Aguirre Moreno, Nuria Andión Rodríguez, Xavier Gómez Guinovart 13

Creación, etiquetación y desambiguación de un corpus de referencia del español
Montserrat Civit Torruella, Irene Castellón Masalles, M. Antonia Martí Antonín 21

Extracción masiva de información sobre subcategorización verbal vasca a partir de corpus
J. Aldazabal, M. Aranzabe, A. Aizua, K. Gojenola, K. Sarasola, Pazzi Goetaga 29

Problemática de la recogida y anotación de una base de datos oral para el gallego
Begoña González Rei, Antonio Cardenal López, Laura Docto Fernández, Carmen García Mateo 37

Gramáticas y formalismos para el análisis morfológico y sintáctico 45

Análisis sintáctico ascendente de TAGs guiado por la esquina izquierda
Vicente Carrillo Montero, Victor J. Díaz Madrigal, Miguel A. Alonso Pardo 47

Una aproximación para resolución de ambigüedad estructural empleando tres mecanismos diferentes
Sofía N. Galicia-Huaro, Alexander F. Gelbukh, Igor A. Bolshakov 55

Aplicaciones industriales del PLN 63

Corpus-based stochastic finite-state predictive text entry for reduced keyboards: application to Catalan
Mikel L. Forcada 65

Integration of dialogue moves and speech recognition in a telephone scenario
José F. Quesada, J. Gabriel Amores, Rafael Ballesteros 71

79

Semántica, pragmática y discurso 79

Dialogue moves for natural command languages
J. Gabriel Amores, José F. Quesada 81

Dialogue management in a home machine environment: linguistic components over an agent architecture
José F. Quesada, Federico García, Esther Sosa, José Angel Bernal, Gabriel Amores 89

Propuesta de un espacio de accesibilidad analítica estructural para textos HTML
Borja Navarro, Patricia Marínuez-Barco, Rafael Muñoz 97

Extracción y recuperación de información 105

Definición de un modelo semántico aplicado a los sistemas de búsqueda de respuestas
José Luis Picado González, Antonio Ferrández Rodríguez 107

Un método de agrupamiento de grafos conceptuales para minería de texto
M. Montes y Gómez, A. Gelbukh, A. López-López, R. Beczar-Yates 115

Normalización de términos multipalabra mediante pares de dependencia sintáctica
Jesús Vilares, Fco. Mario Barcala, Miguel A. Alonso 123

Un modelo de recuperación de información basado en redes bayesianas
Luis M. de Campos, Juan F. Huete, Juan M. Fernández-Luna 131

Generación textual monolingüe y multilingüe 139

WWW como fuente de recursos lingüísticos para su uso en PLN
Fernando Martínez Santiago, L. Alfonso Ureña López, Manuel García Vega 141

Traducción Automática 149

El sistema de traducción automática castellano <-> catalán interNOSTRUM Alacant
R. Canals-Marate, A. Esteve-Guillén, A. Garrido-Atenda, M.J. Guardiola-Savall, Iurraepe-Bellver 151

S. Mouservat-Buendía, S. Ortiz-Rojas, H. Pastor-Pina, P.M. Pérez-Antón, M.L. Forcada 151

MorphTrans: un lenguaje y un compilador para especificar y generar módulos de transferencia morfológica para sistemas de traducción automática
Alicia Garrido-Atenda, Mikel L. Forcada 157

Depósito Legal: B-3941-91
ISSN: 1135-5948

| | |
|---|-----|
| Demostraciones | 307 |
| Análisis y expansión de consultas en lenguaje natural para mejora de la búsqueda en Web <i>Alberto Ruiz, Paloma Martínez, Ana García-Serrano</i> | 305 |
| ANTRO: Un sistema de reconocimiento y gestión de antropónimos <i>Daniel Casanova, Xavier Loré, Rafael Marín, Josep M. Merenciano, Genia Pérez, David Trötzig</i> | 311 |
| Diccionario electrónico de sinónimos y antónimos de la lengua española (DESALE) <i>Santiago Fernández Lanza</i> | 313 |
| EGEO: La estructura geográfica de una base de conocimiento <i>Ignacio González, Sergi Cervell, Josep M. Merenciano, David Trötzig, Joan Vaqué</i> | 315 |
| El uso de editorial de herramientas lingüísticas: un ejemplo con los descriptores humanos <i>Adán Casan, Sergi Cervell, Mireia Colom, Mireia Ferrás, Ignacio González, Rafael Marín, David Trötzig</i> | 317 |
| Gestión flexible de diálogos en el proyecto ADVICE <i>Luis Rodrigo Aguado, Ana García Serrano, Paloma Martínez</i> | 319 |
| Lajji: Un sistema de recuperación multilingüe basado en EuroWordNet <i>Fernando Martínez Santiago, Manuel Carlos Díaz Gallano, L. Alfonso Ureña López,</i> <i>Maite Martín Valdívia, Manuel García Vega, José Ramón Balsa Almagro</i> | 321 |

| | |
|--|-----|
| Lexicografía computacional | 163 |
| Extracción de relaciones léxico-semánticas a partir de palabras derivadas usando patrones de definición <i>Eneko Aguirre, Mikel Lersundi</i> | 165 |
| Etiquetación robusta del lenguaje natural: preprocesamiento y segmentación <i>Jorge Graña Gil, Fco. Mario Barcelona Rodríguez, Jesús Hincas Ferró</i> | 173 |
| Generación automática de familias morfológicas mediante morfología derivativa productiva <i>Jesús Pílares, David Cabrero, Miguel A. Alonso</i> | 181 |
| Internet como fuente de información léxica: extracción de etiquetas de dominio y detección de nuevos sentidos <i>Julio González Arroyo</i> | 185 |
| Resolución de la ambigüedad léxica | 197 |
| A POS-Tagger generator for unknown languages <i>Nuno C. Marques, Gabriel Pereira Lopes</i> | 199 |
| Estudio de cooperación de métodos de desambiguación léxica: marcas de especificidad vs. máxima entropía <i>Aramando Suárez, Andrés Montoyo</i> | 207 |
| Evaluación de un etiquetador morfosintáctico basado en bigramas especializados para el castellano <i>Ferran Pla, Antonio Molina, Natividad Prieto</i> | 215 |

| | |
|---|-----|
| Reconocimiento y síntesis de voz | 223 |
| Asignación automática de marcas de pitch basada en programación dinámica <i>Francisco Alías Pajol, Ignacio Iriondo Sanz</i> | 225 |
| Modelo cuantitativo de entonación del español <i>David Escudero Moncebo, Valentín Cardenosa Payo</i> | 233 |
| Transcriptor ortográfico-fonético para el castellano <i>Maria José Castro, Salvador España, Ismael Salvador, Andrés Marzal</i> | 241 |
| Text to speech --- a rewriting system approach <i>José João Almeida, Alberto Manuel Simões</i> | 247 |
| Una nueva técnica para evaluar sistemas conversacionales basada en la generación automática de diálogos <i>R. López-Cázar, J. C. Segura, A. De la Torre, A. J. Rubio</i> | 255 |
| Análisis automático del contenido textual | 263 |
| Categorización de textos multilingües basada en Redes Neuronales <i>Manuel García Vega, Maite Martín Valdívia, L. Alfonso Ureña López</i> | 265 |
| Cross-lingual keyword assignment <i>Rolf Steinhilber</i> | 273 |
| Generación automática de resúmenes personalizados <i>Ignacio Aveni, Matías Alcázar, Alberto Díaz, José María Gómez, Manuel Mañá</i> | 281 |

| | |
|---|-----|
| Proyectos | 291 |
| Proyecto europeo D'IJonne <i>José F. Quesada, J. Gabriel Amores</i> | 291 |
| XML-It: Procesamientos para gestión de flujo documental multilingüe sobre XML/TEI <i>Joacha Abaitua, Arantza Domínguez, Carmen Izuri, José Luis Ramírez, Inés Jacob</i> | 293 |
| Idioma Modular, Arantza Casillas, Raquel Martínez, Alberto Garay, Thomas Dieckrich..... | 295 |
| Proyecto de indexado automático para documentos en el campo de la física de altas energías <i>Ariano Mantecio Ríos</i> | 297 |
| Proyecto Tagparsing <i>J. Gabriel Amores</i> | 299 |
| Hermes: Servicios de personalización inteligente de noticias mediante la integración de técnicas de análisis automático del contenido textal y modelado de usuario con capacidades bilingües <i>Alfonso Díaz, Manuel de Buenaga, Ignacio Grández, José María Gómez, Antonio García, Inmaculada Chacón, Beatriz San Miguel, Enrique Puertas, Raúl Marcano</i> | 301 |
| Matías Alcázar, Ignacio Aveni, Pablo Cervus..... | 303 |
| Spanish Acquisition: Analysis of learner corpora generated through inter-cultural telecollaboration <i>Julia Kwober, James Lantolf, Steven Thorne, Antonio Jiménez, Brenda Russ, Sagrario Salaberri</i> | 305 |
| Proyecto europeo Siridus <i>José F. Quesada, J. Gabriel Amores</i> | 307 |
| XTRA-BI: Extracción automática de entidades bitextuales para software de traducción asistida <i>Inés Jacob, Joseba Abaitua, Josueta Díaz, José Gómez, Koldo Oñina</i> | 309 |

y también producir una transición del estado de diálogo previo a el(los) siguiente(s). Finalmente, los actos de habla también contienen otro tipo de información: los detalles del diálogo (nombres, datos, características, etc.). Esta información es estática para toda la sesión y se almacena en el modelo de sesión.

El componente que se encarga de la gestión del diálogo tiene el control de la conversación (los aspectos formales a través de los estados del diálogo y los aspectos intencionales a través de los hilos). Después de procesar la información que contienen los actos de habla, el generador de discurso debería ser capaz de construir una respuesta del sistema. En este punto puede ser necesario cierto tipo de conocimiento para completar la respuesta. Este conocimiento puede encontrarse en el contexto del diálogo, o bien, si se trata de conocimiento nuevo el agente de interacción se lo pedirá al agente inteligente. La respuesta del agente inteligente es una oferta configurable modelizada como un árbol de decisión. El generador de discurso extraerá del árbol la información necesaria para alcanzar una solución, requiriendo ciertas explicaciones al usuario, si fuesen necesarias. Una vez se tiene toda la información necesaria se construye la estructura semántica, compuesta por una serie de actos de habla enlazados. El agente de interfaz será el encargado de traducir esos *speech acts* en frases, ítems, menús, fotografías, movimientos del avatar 3D, o una combinación de ellas.

Referencias

- Cohen, P.R., Levesque, H.J. (1991) *Confirmation and Joint Action*, Proceedings of International Joint Conf. on Artificial Intelligence, 1991
- Cole, R.A. ed. (1997) *Survey of the state of the art in Human Language Technology*.
- Cooper, R., Larsson, S., (1999) *Dialogue moves and information states*. In Proc. of the Third IWCS, Tilburg, 1999.
- Seale, J.R., (1969) *Speech Acts: an essay in the philosophy of language*. Cambridge Univ. Press.

| | |
|----------------------------|--|
| Representative acts | [i]: type (confirmation / data / ...) [m]: manner (approval/objection/...) [s]: subject (product/verb/system/...) [c]: content (...) |
| Inform: < m, s, c > | Example: / m, d, i, inform(data, identity, user, AD) |
| Authoritative acts | [m]: manner (start / offer / task / ...) [a]: allowable (open / closed) |
| Authorize: < m, a > | Example: Com / help you? authorize(task, open) |
| Directive acts | [i]: type (information/comparison/...) [m]: manner (approval/objection/...) [c]: subject (user/system/...) [c]: content (values...) |
| Request: < i, m, s, c > | Examples: Who are...?, request(data, identity, ...) |
| Command: < i, s, c > | Show me some more: command(search, system, product) |
| Null Speech | Null: < > |

Figura 1. Conjunto de actos de habla del proyecto ADVICE

El sistema ADVICE contempla dos tipos de entrada de datos: en lenguaje natural (en inglés) a través del módulo intérprete de lenguaje, y a través del interfaz gráfico de usuario (navegando a través de las opciones de la página Web). El agente de interfaz se encarga de transformar las diferentes entradas del usuario a estructuras semánticas (evidencias de actos de habla) y se los pasa al agente de interacción (la parte del sistema encargada de la gestión del diálogo). Para presentar datos al usuario, existen tres posibilidades: el generador de LN, el interfaz gráfico de usuario y un avatar en 3D. Las estructuras semánticas de salida que genera el agente de interacción (actos de habla también) son la entrada para el agente de interfaz, que se encarga de construir una salida coordinada aprovechando los diferentes canales antes mencionados.

Las intervenciones en el diálogo se dividirán en 'piezas de discurso' (subconjunto del discurso con significado independiente), y cada uno de ellos es representado como un conjunto de actos de habla.

A lo largo de una intervención, hay que adaptar y actualizar el hilo o línea locutiva

L.Lajú: Un sistema de recuperación multilingüe basado en EuroWordNet

Fernando Martínez Santiago

Manuel Carlos Díaz Galiano

L. Alfonso Ureña López

Maite Martín Valdivia

Manuel García Vega

José Ramón Balsas Almagro

Departamento de Informática, Universidad de Jaén, Spain

{dofer, mediaz, lauren, maite, mgarcia, jrbalsas}@ujaen.es

Resumen Se presenta aquí un sistema de recuperación de información multilingüe completamente funcional, capaz de procesar consultas en inglés y español, recuperando indistintamente documentos en ambos idiomas.

1 Introducción

El sistema de recuperación de información multilingüe L.Lajú aquí presentado viene a ser una implementación del modelo sugerido en [1], con la mejora en el cálculo de probabilidades de traducción a partir de SemCor tal como se describe en [2]. El sistema de recuperación que proponemos es capaz de recuperar artículos de las secciones de nacional e internacional de los sitios de "ABC", "El Mundo" y "El País", y de las secciones de internacional del "Washington Post", "CNN news" y "The Guardian Observer", correspondientes al año 2001. El proceso de recopilación y formateado del corpus se ha llevado a cabo usando la herramienta WebReader[3].

2 Descripción de Lajú

L.Lajú consta de tres partes bien diferenciadas:

- i. un motor de búsqueda multilingüe
- ii. una interfaz basada en el Web
- iii. una utilidad para la adquisición de nuevos documentos

El motor de búsqueda

El motor de búsqueda requiere un preprocesamiento, que pretende resolver las barreras lingüísticas existentes, para posteriormente indexar los documentos con independencia del lenguaje origen, usando para ello el sistema IR ZPrise. La elección de ZPrise ha venido determinada por su disponibilidad y por tratarse de un sistema recomendado en tareas CLIR como la que aquí se presenta. En cualquier caso, antes de indexar un documento, se requiere cierto preprocesamiento que a continuación describimos:

- i. Detección de multi-palabras registradas en EuroWordNet.
- ii. Extracción del lema para cada término que no forme parte de una multipalabra. Si se trata de un documento en español se ha usado MACO+RELAX [4]. En caso de tratarse de un término inglés, se ha optado por el Brill Tagger, junto con el lematizador que se encuentra en WordNet.
- iii. Resolución de la ambigüedad léxica. Para esta tarea, se ha usado el desambiguador propuesto en [5].
- iv. Obtención del *synset* al que pertenece el término ya desambiguado. De esta manera, conseguimos dos ventajas: por una parte estamos indexando por concepto antes que por términos, y además conseguimos un índice independiente de lenguaje.

