

Artículos

Aspectos ortográficos, léxicos y morfosintácticos del etiquetado lingüístico de un corpus de informática en lengua gallega <i>José Luis Aguirre Moreno, Nuria Andión Rodríguez, Xavier Gómez Guinovart</i>	13
Creación, etiquetación y desambiguación de un corpus de referencia del español <i>Montserrat Civit Torruella, Irene Castellón Masalles, M. Antònia Martí Antonin</i>	21
Extracción masiva de información sobre subcategorización verbal vasca a partir de corpus <i>I. Aldezabal, M. Aranzabe, A. Atutxa, K. Gojenola, K. Sarasola, Patxi Goenaga</i>	29
Problemática de la recogida y anotación de una base de datos oral para el gallego <i>Begoña González Rei, Antonio Cardenal López, Laura Docío Fernández, Carmen García Mateo</i>	37
Análisis sintáctico ascendente de TAGs guiado por la esquina izquierda <i>Vicente Carrillo Montero, Víctor J. Díaz Madrigal, Miguel A. Alonso Pardo</i>	47
Una aproximación para resolución de ambigüedad estructural empleando tres mecanismos diferentes <i>Sofía N. Galicia-Haro, Alexander F. Gelbukh, Igor A. Bolshakov</i>	55
Corpus-based stochastic finite-state predictive text entry for reduced keyboards: application to Catalan <i>Mikel L. Forcada</i>	65
Integration of dialogue moves and speech recognition in a telephone scenario <i>José F. Quesada, J. Gabriel Amores, Rafael Ballesteros</i>	71
Dialogue moves for natural command languages <i>J. Gabriel Amores, José F. Quesada</i>	81
Dialogue management in a home machine environment: linguistic components over an agent architecture <i>José F. Quesada, Federico García, Esther Sena, José Angel Bernal, Gabriel Amores</i>	89
Propuesta de un espacio de accesibilidad anafórica estructural para textos HTML <i>Borja Navarro, Patricio Martínez-Barco, Rafael Muñoz</i>	97
Definición de un modelo semántico aplicado a los sistemas de búsqueda de respuestas <i>José Luis Vicedo González, Antonio Ferrández Rodríguez</i>	107
Un método de agrupamiento de grafos conceptuales para minería de texto <i>M. Montes-y-Gómez, A. Gelbukh, A. López-López, R. Baeza-Yates</i>	115
Normalización de términos multipalabra mediante pares de dependencia sintáctica <i>Jesús Vilares, Fco. Mario Barcala, Miguel A. Alonso</i>	123
Un modelo de recuperación de información basado en redes bayesianas <i>Luis M. de Campos, Juan F. Huete, Juan M. Fernández-Luna</i>	131
WWW como fuente de recursos lingüísticos para su uso en PLN <i>Fernando Martínez Santiago, L. Alfonso Ureña López, Manuel García Vega</i>	141
El sistema de traducción automática castellano <-> catalán interNOSTRUM Alacant <i>R. Canals-Marote, A. Esteve-Guillén, A. Garrido-Alenda, M.I. Guardiola-Savall, J. Iturraspe-Bellver, S. Montserrat-Buendía, S. Ortiz-Rojas, H. Pastor-Pina, P.M. Pérez-Antón, M.L. Forcada</i>	151
MorphTrans: un lenguaje y un compilador para especificar y generar módulos de transferencia morfológica para sistemas de traducción automática <i>Alicia Garrido-Alenda, Mikel L. Forcada</i>	157
Extracción de relaciones léxico-semánticas a partir de palabras derivadas usando patrones de definición <i>Eneko Agirre, Mikel Lersundi</i>	165
Etiquetación robusta del lenguaje natural: preprocesamiento y segmentación <i>Jorge Graña Gil, Fco. Mario Barcala Rodríguez, Jesús Vilares Ferro</i>	173
Generación automática de familias morfológicas mediante morfología derivativa productiva <i>Jesús Vilares, David Cabrero, Miguel A. Alonso</i>	181
Internet como fuente de información léxica: extracción de etiquetas de dominio y detección de nuevos sentidos <i>Celina Santamaría, Julio Gonzalo Arroyo</i>	189
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>Armando 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
Asignación automática de marcas de pitch basada en programación dinámica <i>Francesc Alías Pujol, Ignasi Iriando Sanz</i>	225
Modelo cuantitativo de entonación del español <i>David Escudero Mancebo, Valentín Cardenoso Payo</i>	233
Transcriptor ortográfico-fonético para el castellano <i>María José Castro, Salvador España, Ismael Salvador, Andrés Marzal</i>	241



MINISTERIO
DE CIENCIA
Y TECNOLOGÍA

SEPLN



JUNTA DE ANDALUCÍA
Consejería de Educación y Ciencia



DIPUTACIÓN PROVINCIAL
DE JAÉN



KIBUTZ.COM

a2 INFORMATICA



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-Cózar, J. C. Segura, A. De la Torre, A. J. Rubio</i>	255
Categorización de textos multilingües basada en Redes Neuronales <i>Manuel García Vega, Maite Martín Valdivia, L. Alfonso Ureña López</i>	265
Cross-lingual keyword assignment <i>Ralf Steinberger</i>	273
Generación automática de resúmenes personalizados. <i>Ignacio Acero, Matías Alcojor, Alberto Díaz, José María Gómez, Manuel Maña</i>	281
Proyectos	
Proyecto europeo D'Homme <i>José F. Quesada, J. Gabriel Amores</i>	291
XML-Bi: Procesamientos para gestión de flujo documental multilingüe sobre XML/TEI <i>Joseba Abaitua, Arantza Domínguez, Carmen Isasi, José Luis Ramírez, Inés Jacob, Idoia Madariaga, Arantza Casillas, Raquel Martínez, Alberto Garay, Thomas Diedrich</i>	293
Proyecto de indexado automático para documentos en el campo de la física de altas energías <i>Arturo Montejó Ráez</i>	295
Proyecto Tagparsing <i>J. Gabriel Amores</i>	297
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 <i>Alberto Díaz, Manuel de Buenaga, Ignacio Giráldez, José María Gómez, Antonio García, Inmaculada Chacón, Beatriz San Miguel, Enrique Puertas, Raúl Murciano, Matías Alcojor, Ignacio Acero, Pablo Gervás</i>	299
Spanish Acquisition: Analysis of learner corpora generated through inter-cultural telecollaboration <i>Julia Kasher, James Lantolf, Steven Thorne, Antonio Jiménez, Brenda Ross, Sagrario Salaberri</i>	301
Proyecto europeo Siridus <i>José F. Quesada, J. Gabriel Amores</i>	303
XTRA-Bi: Extracción automática de entidades bitextuales para software de traducción asistida <i>Inés Jacob, Joseba Abaitua, Josuka Díaz, Josu Gómez, Koldo Ocina</i>	305
Demostraciones	
Análisis y expansión de consultas en lenguaje natural para mejora de la búsqueda en Web <i>Alberto Ruiz, Paloma Martínez, Ana García-Serrano</i>	309
ANTRO: Un sistema de reconocimiento y gestión de antropónimos <i>Daniel Casanova, Xavier Lloré, Rafael Marín, Josep M. Merenciano, Gema Pérez, David Trotzig</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 Trotzig, Joan Vaqué</i>	315
El uso de editorial de herramientas lingüísticas: un ejemplo con los descriptores humanos <i>Adán Cassan, Sergi Cervell, Mireia Colom, Mireia Farrís, Ignacio González, Rafael Marín, David Trotzig</i>	317
Gestión flexible de diálogos en el proyecto ADVICE <i>Luis Rodrigo Aguado, Ana García Serrano, Paloma Martínez</i>	319
Llajú: Un sistema de recuperación multilingüe basado en EuroWordNet <i>Fernando Martínez Santiago, Manuel Carlos Díaz Galiano, L. Alfonso Ureña López, Maite Martín Valdivia, Manuel García Vega, Jose Ramón Balsas Almagro</i>	321

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 Buenaga (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 Galiano (Universidad de Jaén)

José Ramón Balsas (Universidad de Jaén)

Pedro González García (Universidad de Jaén)

Víctor Rivas Santos (Universidad de Jaén)

REVISORES EXTERNOS:

Eneko Agirre Bengoa
Guadalupe Aguado
Pablo Aibar Ausina
Iñaki Alegria Loinaz
Manuel Alonso González
Miguel A. Alonso Pardo
Margarita Alonso Ramos
Alberto Álvarez Ladrón
Montserrat Arévalo Rodríguez
María Victoria Arranz Corzana
David Cabrero Souto
José Carlos González
Juan Carlos Pérez
Xavier Carreras Pérez
Núria Castell Ariño
María José Castro Bleda
Montserrat Civit Torruella
Salvador Climent Roca
Arantza Díaz de Ilarraza
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

Àngels Hernández Gómez
Eduardo Lleida Solano
Joaquín Llisteri
Fernando Llopis
Lluís Màrquez Villodre
M. Antonia Martí Antonín
Paloma Martínez Fernández
Andrés Montoyo Guijarro
Antonio Moreno Sandoval
Javier Pérez Guerra
Ferran Pla Santamaría
Celia Rico Pérez
German Rigau
Fernando Sáenz Pérez
Fernando Sánchez León
Antonio Sánchez Valderrábanos
Emilio Sanchis Arnal
Encarna Segarra Soriano
María José Simón Aragón
Alejandro Sobrino Cerdeiriña
Armando Suárez Cueto
Declerck Thierry
Antonio Valderrábanos
Gloria Vázquez
Julio Villena Román
Jorge Vivaldi

Procesamiento del

SEPLN

Artículos

Lingüística del cor
Aspectos ortográficos
de informática en len
José Luis Aguirre M
Creación, etiquetaci
Montserrat Civit Tor
Extracción masiva de
I. Aldezabal, M. Aran
Problemática de la res
Begoña González Rei

Gramáticas y forma
Análisis sintáctico asc
Vicente Carrillo Moni
Una aproximación pa
Sofía N. Galicia-Haro

Aplicaciones industr
Corpus-based stochast
Mikel L. Forcada.....
Integration of dialogue
José F. Quesada, J. Ga

Semántica, pragmática
Dialogue moves for nat
J. Gabriel Amores, Jos
Dialogue management
José F. Quesada, Feder
-Propuesta de un espaci
Borja Navarro, Patrici

Extracción y recupera
- Definición de un model
José Luis Vicedo Gonzá
Un método de agrupami
M. Montes-y-Gómez, A.
Normalización de térmi
Jesús Vilares, Fco. Mari
- Un modelo de recuperac
Luis M. de Campos, Juan

Generación textual mo
- WWW como fuente de re
Fernando Martínez Sant

Traducción Automática
El sistema de traducción
R. Canals-Marote, A. Est
S. Montserrat-Buendia, S
MorphTrans: un lenguaje
de transferencia morfológ
Alicia Garrido-Alenda, M

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

Artículos

Lingüística del corpus	11
Aspectos ortográficos, léxicos y morfosintácticos del etiquetado lingüístico de un corpus de informática en lengua gallega <i>José Luis Aguirre Moreno, Nuria Andión Rodríguez, Xavier Gómez Guinovart</i>	13
Creación, etiquetación y desambiguación de un corpus de referencia del español <i>Montserrat Civit Torruella, Irene Castellón Masalles, M. Antònia Martí Antonin</i>	21
Extracción masiva de información sobre subcategorización verbal vasca a partir de corpus <i>I. Aldezabal, M. Aranzabe, A. Atutxa, K. Gojenola, K. Sarasola, Patxi Goenaga</i>	29
Problemática de la recogida y anotación de una base de datos oral para el gallego <i>Begoña González Rei, Antonio Cardenal López, Laura Docío Fernández, Carmen García Mateo</i>	37
Gramáticas y formalismos para el análisis morfológico y sintáctico	45
Análisis sintáctico ascendente de TAGs guiado por la esquina izquierda <i>Vicente Carrillo Montero, Víctor J. Díaz Madrigal, Miguel A. Alonso Pardo</i>	47
Una aproximación para resolución de ambigüedad estructural empleando tres mecanismos diferentes <i>Sofía N. Galicia-Haro, Alexander F. Gelbukh, Igor A. Bolshakov</i>	55
Aplicaciones industriales del PLN	63
Corpus-based stochastic finite-state predictive text entry for reduced keyboards: application to Catalan <i>Mikel L. Forcada</i>	65
Integration of dialogue moves and speech recognition in a telephone scenario <i>José F. Quesada, J. Gabriel Amores, Rafael Ballesteros</i>	71
Semántica, pragmática y discurso	79
Dialogue moves for natural command languages <i>J. Gabriel Amores, José F. Quesada</i>	81
Dialogue management in a home machine environment: linguistic components over an agent architecture <i>José F. Quesada, Federico García, Esther Sena, José Ángel Bernal, Gabriel Amores</i>	89
Propuesta de un espacio de accesibilidad anafórica estructural para textos HTML <i>Borja Navarro, Patricio Martínez-Barco, Rafael Muñoz</i>	97
Extracción y recuperación de información	105
Definición de un modelo semántico aplicado a los sistemas de búsqueda de respuestas <i>José Luis Vicedo González, Antonio Ferrández Rodríguez</i>	107
Un método de agrupamiento de grafos conceptuales para minería de texto <i>M. Montes-y-Gómez, A. Gelbukh, A. López-López, R. Baeza-Yates</i>	115
Normalización de términos multpalabra mediante pares de dependencia sintáctica <i>Jesús Vilares, Fco. Mario Barcala, Miguel A. Alonso</i>	123
Un modelo de recuperación de información basado en redes bayesianas <i>Luis M. de Campos, Juan F. Huete, Juan M. Fernández-Luna</i>	131
Generación textual monolingüe y multilingüe	139
WWW como fuente de recursos lingüísticos para su uso en PLN <i>Fernando Martínez Santiago, L. Alfonso Ureña López, Manuel García Vega</i>	141
Traducción Automática	149
El sistema de traducción automática castellano <-> catalán interNOSTRUM Alacant <i>R. Canals-Marote, A. Esteve-Guillén, A. Garrido-Alenda, M.I. Guardiola-Savall, Iturraspe-Bellver, S. Montserrat-Buendia, S. Ortiz-Rojas, H. Pastor-Pina, P.M. Pérez-Antón, M.L. Forcada</i>	151
MorphTrans: un lenguaje y un compilador para especificar y generar módulos de transferencia morfológica para sistemas de traducción automática <i>Alicia Garrido-Alenda, Mikel L. Forcada</i>	157

163	Demostraciones.....	
165	Análisis y expansión de consultas en lenguaje natural para mejora de la búsqueda en Web <i>Alberto Ruiz, Paloma Martínez, Ana García-Serrano</i>	
173	ANTRO: Un sistema de reconocimiento y gestión de antropónimos <i>Daniel Casanova, Xavier Lloré, Rafael Marín, Josep M. Merenciano, Gema Pérez, David Trotzig</i>	311
181	Diccionario electrónico de sinónimos y antónimos de la lengua española (DESALE) <i>Santiago Fernández Lanza.....</i>	313
185	EGEO: La estructura geográfica de una base de conocimiento <i>Ignacio González, Sergi Cervell, Josep M. Merenciano, David Trotzig, Joan Vaqué.....</i>	315
197	El uso de editorial de herramientas lingüísticas: un ejemplo con los descriptores humanos <i>Adán Cassan, Sergi Cervell, Mireia Colom, Mireia Farrís, Ignacio González, Rafael Marín, David Trotzig</i>	317
199	Gestión flexible de diálogos en el proyecto ADVICE <i>Luis Rodrigo Aguado, Ana García Serrano, Paloma Martínez.....</i>	319
207	Llajú: Un sistema de recuperación multilingüe basado en EuroWordNet <i>Fernando Martínez Santiago, Manuel Carlos Díaz Galiano, L. Alfonso Ureña López, Maite Martín Valdivia, Manuel García Vega, Jose Ramón Balsas Almagro.....</i>	321
215		
223		
225		
233		
241		
247		
255		
263		
265		
273		
281		
291		
293		
295		
297		
299		
301		
303		
305		

Proyecto de indexado automático para documentos en el campo de la Física de Altas Energías

Arturo Montejo Ráez
Laboratorio Europeo de Física de Partículas (CERN)
Ginebra (Suíza)

Resumen Se describe aquí el sistema *HEPindexer*, un indexador automático para documentos sobre Física de Altas Energías. En su primera fase se ha conseguido la proposición de palabras clave primarias usando el tesoro del laboratorio alemán DESY. Los resultados, utilizando un enfoque estadístico, esperan la consecución de una herramienta eficaz de ayuda en el proceso de indexado.

Palabras clave: *indexación automática, modelo de espacio vectorial, modelo estadístico, tesoro.*

1 Introducción

Este proyecto consiste en el desarrollo de un sistema automático de indexado por asignación. El indexado por asignación consiste en la selección de palabras clave dentro de un léxico controlado (en nuestro caso un tesoro) que describan y resuman los conceptos más importantes tratados en un texto dado. El sistema propone palabras clave según el tesoro del laboratorio alemán DESY (Deutsche Elektronen-Synchrotron) a partir de artículos completos en inglés relacionados con Física de Altas Energías.

El sistema implementado toma documentos en los formatos PDF, PostScript y texto plano (ASCII) y genera una lista de palabras clave *primarias*, pues en su primera fase no se consideran las palabras clave complementarias (o *secundarias*).

2 Trabajos anteriores

El indexado automático basado en la frecuencia de palabra se remonta a los años 50 y los trabajos de Luhn [4] y Baxendale [1]. Posteriormente han aparecido otros muchos trabajos y algunos sistemas de indexado por asignación integrados como sistemas de ayuda. Entre estos sistemas podemos citar

BIOSIS de Vleduts-Stokolov [8], el sistema MeSH de ayuda al indexado de la Biblioteca Nacional de Medicina Estadounidense [5] y, sobre todo, el NASA MAI System [3]. Éste último es uno de los sistemas que mayor éxito han tenido y cuyo uso viene a demostrar la efectividad de la ayuda automática en el proceso de indexado.

3 El problema del indexado

El servidor *weblib.cern.ch* cubre más de 430,000 referencias bibliográficas y 170,000 documentos en formato electrónico, relacionados con el CERN y con la Física de Altas Energías. El uso de palabras clave para su clasificación y búsqueda es muy importante, siendo el laboratorio de DESY el encargado de etiquetar los documentos gracias a su equipo de indexadores (ver ejemplo en figura 1). Pero el creciente

electron positron, colliding beam
K*(892), hadronic decay
mass spectrum, (K0(S) pi-)
antimatter
experimental results
electron positron --> tau+ tau-
K*(892) --> K- pi0
10.6 GeV-cms

Figura 1: Ejemplo de palabras clave propuestas por DESY para un artículo.

volumen de artículos hace que el esfuerzo empleado se vea desbordado. Para ello un sistema de ayuda a la indexación manual es requerido en este entorno.

4 Descripción del sistema

El modelo utilizado sigue el modelo de espacio vectorial propuesto por Salton [7]. Los textos se procesan aplicando eliminación de palabras vacías [2], *stemming* [6] y limitando

el vector a los términos con mayor peso según su frecuencia y su frecuencia inversa de documento (*idf*). A partir de la colección de entrenamiento, en la cual se dispone del texto íntegro asociado a las claves DESY, se calculan, usando métodos estadísticos, las siguientes relaciones:

Relación entre un término y un documento. Corresponde con el vector del texto íntegro.

Relación entre una palabra clave y un documento. Se obtiene a partir de la colección de entrenamiento. Es, básicamente, la frecuencia de la clave primaria en las claves DESY del documento.

Relación entre una palabra clave y un término. Usando los documentos como nexo de unión, se calcula el valor del peso entre cada clave y cada término. Este valor se normaliza haciendo uso de lo que hemos denominado *frecuencia inversa de palabra clave* (IKF) que penaliza aquellos términos que se relacionan con muchas palabras clave.

Esta última relación es la matriz que se usa para calcular un *ranking* de claves a partir de un nuevo documento de texto. El sistema propone las claves con mayor peso como posibles claves primarias para el texto en cuestión.

5 Resultados

El sistema se entrenó a partir de una colección de 2,400 documentos y los primeros tests se realizaron sobre una colección de 1,200 documentos. Los resultados muy satisfactorios (actualmente 53,8% en precisión y 59,7% cobertura, ver figura 2).

6 Conclusión

Su interfaz web permite el amplio uso de la herramienta la cual, aún en su primera fase, promete convertirse en una valiosa ayuda a la indexación realizada manualmente. Actualmente el sistema se encuentra en producción en el servidor de documentos del CERN.

Este sistema demuestra que un algoritmo estadístico sencillo puede ser usado en un entorno con vocabulario técnico, donde la ambigüedad es baja dada la especialización del área.

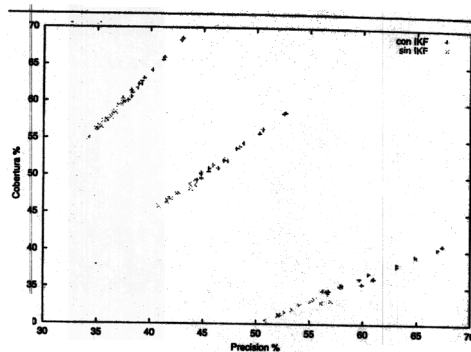


Figura 2: Influencia de IKF sobre la precisión y la cobertura

7 Trabajo futuro

Aún quedan problemas por resolver, como la generación de claves sobre desintegración de partículas y otros problemas que sólo aparecen en el ámbito de la Física de Altas Energías. Se pretende en futuras fases el uso de recursos lingüísticos (como WordNet), así como el estudio del algoritmo para proponer claves secundarias. Este sistema ya ha despertado el interés de expertos de la NASA trabajando con el actual NASA MAI System.

Referencias

- [1] P. Baxendale. Machine-made index for technical literature — an experiment, 1958.
- [2] William B. Frakes and Ricardo Baeza-Yates. Information retrieval: Data structures and algorithms, 1992.
- [3] Paul H. Klingbie June P. Silvester, Michael T. Genuardi. Machine-aided indexing at nasa, 1994.
- [4] H. Luhn. A statistical approach to mechanized encoding and searching of literary information, 1957.
- [5] National Library of Medicine, Bethesda, US. *Medical Subject Headings (MeSH)*, 1993.
- [6] M.F. Porter. An algorithm for suffix stripping, 1980.
- [7] G. Salton. A vector space model for automatic indexing, 1975.
- [8] Natasha Vieduts-Stokolo. Concept recognition in an automatic text-processing system for the life sciences, 1987.

1 Ficha del

- Título del técnicas y he ing y unifica
- Entidad Final de Investi Ministerio d porte. (Cód Diciembre 1
- Grupos Par Sevilla y Un
- Investigador Amores. D glesa. Uni de la Fronte 95 455 1549 Toni Badia. Fabra. tbad

2 Resumen

La integración cionales para el una prioridad (tuales de investi po de la Ingenie Dentro del c Lenguaje Natur ciones requierer guna de sus fa parsing y unifi ciones, merecen de Información la búsqueda de temas de gestión voz, traducción Durante los pos participant tificación en el ramientas para putacional (U técnicas de pa samiento del L