Jornada de Seguimiento de Proyectos, 2009
Programa Nacional de Tecnologías Informáticas

TEXT-MESS: Intelligent, Interactive and Multilingual Text Mining based on Human Language Technologies TIN2006-15265-C06

Patricio Martínez-Barco[*]
Coordinador, U. de Alicante

Julio Gonzalo[†]
IP, UNED

L. Alfonso Ureña[‡]
IP, U. de Jaén

Ferran Pla[§]
IP, U. Politécnica de Valencia

Alicia Ageno[**]
IP, U. Politécnica de Cataluña

M. Antònia Marti[††]
IP, U. de Barcelona

Abstract

The goal of the project is to analyze, experiment, and develop intelligent, interactive and multilingual Text Mining technologies, as a key element of the next generation of search engines, systems with the capacity to find "the need behind the query". This new generation will provide specialized services and interfaces according to the search domain and type of information needed. Moreover, it will integrate textual search (websites) and multimedia search (images, audio, video), it will be able to find and organize information, rather than generating ranked lists of websites.
Keywords: Text Mining, Human Language Technologies (HLT), HLT resources, Information Retrieval, Question Answering, Information Extraction, HLT evaluation

1 Project Goals

The main goal of this project is to develop intelligent, interactive, multilingual Text Mining technologies, in which documental search on Web pages, multimedia search on images and search on partially-structured documents are integrated and based on HLT. To carry out this goal, we propose three basic lines: **(G1)** The study and development of Text Mining systems (search, extraction, classification, information retrieval and text analysis from a Human Language Technology -HLT- perspective), analysing on the one hand interactive, multilingual aspects, with special emphasis into Catalan and Spanish, and, on the other, the efficiency and effectiveness of these systems in written documents, oral transcripts and images both in open domains (namely, the web) and restricted ones (biomedicine and tourism). **(G2)** The adaptation and improvement of previously existing resources and tools (greater coverage, better quality and treatment of restricted domains) and creation of new tools required to undertake new applications based on HLT, by

[*] (UA) (TIN2006-15265-C06-01) Email: patricio@dlsi.ua.es
[†] (UNED) (TIN2006-15265-C06-02) Email: julio@lsi.uned.es
[‡] (UJA) (TIN2006-15265-C06-03) Email: laurena@ujaen.es
[§] (UPV) (TIN2006-15265-C06-04) Email: fpla@dsic.upv.es
[**] (UPC) (TIN2006-15265-C06-05) Email: ageno@lsi.upc.edu
[††] (UB) (TIN2006-15265-C06-06) Email: amarti@ub.edu

combining linguistic knowledge and Machine Learning (ML) techniques. **(G3)** The connection of this research with the main international evaluation campaigns on search systems and HLT. On the one hand, involved groups will participate in these campaigns with the aim of contrasting their research results with those of the main international groups; on the other, we will promote and coordinate some of the tasks, with the goal of promoting the research in the interest lines of the project.

To achieve these goals, the project has been organized into five modules. Moreover, each module has been divided into activities and tasks. A temporal scheduling of activities in the project with their coordinators is shown in Figure 1

MOD / ACT		Coordinator MOD	Coordinator ACT	YEAR 1				YEAR 2				YEAR 3			
M1	A1	P. Martínez (UA)													
M2	A1	F.Pla (UPV)	T. Martí (UB)												
	A2		A. Ageno (UPC)												
	A3		F. Martínez (UJA)												
	A4		E. Sanchís (UPV)												
	A5		J. Turmo (UPC)												
M3	A1	A. Ureña (UJA)	H. Rodríguez (UPC)												
	A2		A. Molina (UPV)												
	A3		A. Ferrández (UA)												
	A4		M. Martín (UJA)												
M4	A1	J. Gonzalo (UNED)	A. Peñas (UNED)												
	A2		A. Montoyo (UA)												
	A3		J. Gonzalo (UNED)												
M5	A1	A. Suárez (UA)													

Figure 1. Chronogram of activities

2 Level of achievement reached in the Project

The project has been carried out on time and on budget. This, added to the attainment of the planned objectives, shows that project development and management is satisfactory. Next, we are going to describe the state of the Project in relation with the objectives and modules planed, as well as the main scientific and technological results.

Module 1 - Project management and coordination

Five meetings have been organized during the project. Alicante, Feb. 2007 (kick-off); Barcelona, Jun. 2007 (coordination and monitoring); Barcelona, Oct. 2007 (plenary meeting); Jaén, May. 2008 (coordination and monitoring) and Valencia, Oct. 2008 (status review and new project submission planning). Moreover, during the meeting in october 2007, the 1st. TEXT-MESS Seminar was held where the involved teams showed their results. Several email lists were created according to activities and modules, and a collaborative wiki was developed in http://gplsi.dlsi.ua.es/text-mess. During the first semester of the project, due to unexpectedly new duties of the project coordinator (Manuel Palomar), a change on this figure to Patricio Martínez-Barco was carried out with the approval of the *Dirección General de Investigación*. Then, a new distribution of activities and task coordinators were developed conforming the final configuration showed in Figure 1.

Module 2 - Human Language Technology resources, methods, techniques and tools oriented to intelligent text mining

This module is oriented to the survey, adaptation and development of resources and basic tools to support the generation of ambitius subsystems of information extraction, information retrieval or

question answering, achieving a wider coverage, and improving the quality and the processing of specific domains. To acomplish this module, five activities where planified.

M2A1 - Development of resources and methods.

For Romanic languages, particularly for Spanish and Catalan, a competitive quality and coverage level have been achieved both at the morphosyntactic and shallow syntactic level. The goal of this activity is to achieve equivalent levels to those achieved for English at sentence semantics and deep syntactic analysis.

Results: Resources for ambiguity resolution at the deep syntactic and semantic level. The development of an annotation guide and the methodology to build corpora with this information was published in [67] and [69]. Also, an annotation scheme for the anaphoric annotation of multilingual Question Answering corpora has been published in [40]. **Assigning thematic roles to syntactic functions** (semantic analysis of predicates). A method to automatic role labeling based on ML from corpora with thematic roles related to syntactic functions for each verb sense has been developed in [68]. Methodology combination strategies in order to increase robustness and to gain coverage and independence from parse errors were developed [196] [204]. **Annotated corpus completion for testbed of quality control of the developed tools, applications and techniques.** A Spanish corpus with 6000 questions classified into 50 different categories has been collected to train question classification modules to QA tasks [39]. Two corpora of 500.000 words each – AnCora-Ca for Catalan and AnCora-ES for Spanish- with morphological [231], syntactic (constituents and functions, [224]), semantic (Named Entities [227], [236]; WordNet synsets; Thematic Roles[228], [232]; semantic classes [245]) and pragmatic information (coreference, [230], [240]) had been developed. These corpora are freely available at http://clic.ub.edu/ancora, [244], [225], [226]. CesCa corpus. that consists of 2.400 texts written by catalan scholars of primary and secondary school. (http://clic.ub.edu/cesca); GeoSemCor, for the toponyms disambiguation task [165][133]; Emoticorpus, for automatic recognition of humor [147]; Geo-WordNet, for geographical information [161][165][166]; ANERcorp and ANERgazet, for named entities recognition and gazetteers, respectively, for the Arabic language. All these resources can be downloaded from: http://www.dsic.upv.es/grupos/nle/downloads.html. Two verbal lexions, AnCoraVerb-ES for Spanish and AnCoraVerb-CA for Catalan, have been developed. These lexicons, of about 2.000 entries, contain information of the constituent structure, thematic roles and semantic class of each verb. [233], [234]. Development of Dial-Cat and Histo-Cat corpora morphologically annotated (http://stel.ub.edu/dialcat, http://stel.ub.edu/histocat) [247]. **Automatic construction and updating of specific domain ontologies.** The project is working on a pharmacological ontology and its relationship with the MultiWordNet resource. Also, we have developed an ontology oriented to Question Answering systems applied to medical domains, from UMLS and MultiWordNet resources [37] [33]. Development of the Arabic Wordnet [242]. **Global platform.** A global platform to integrate Natural Language Processing Resources and Tools, called InTIME, has been designed and created [8].

M2A2 - Development of tools.

This activity captures several proposals for the creation of new basic tools which essential for intelligent text mining in the languages involved in the project. These tools will be created from the resources, methods and corpora obtained as a result of the previous task and also from already existing resources.

Results: Document clustering. A new method to web page clustering based on name disambiguation was published in [49]. Besides, non-supervised document clustering methods based on the combination of simple clustering methods have been explored. In particular, two possible

strategies have been presented in [197]. Moreover, a work on Text Classification using Support Vector Machines is in progress [205]. A method using Fuzzy Logic Based Representation was published in [87]. Development of tools for clustering narrow-domain short texts are [135][140][147][157][164][168]. **Pattern acquisition for Information Extraction.** We have been working in the integration of semantics into non-supervised learning of relationships. Moreover, we have been working in the acquisition of temporal information extraction patterns, a work that has led to publication [206]. **Automatic learning of paraphrases.** Some techniques to paraphrases detection based on textual entailment have been tested in [40]. Furthermore, the interface (COCO, http://www.lsi.upc.edu/~textmess/index.php TextMess Corpora Compilation) was developed in order to acquire material to machine learning paraphrase. Currently Catalan, Spanish, English and Arabic are the languages supported. **Construction of a tool for Recognition and Classification of Weak Named Entities (NERC).** Several experiments on the combination of ML methods have been developed [47] [50], and context influences on the categorization and discrimination of person names on Spanish and Portuguese [19]. In [17][50] the extension of NERC tool to new languages like Italian and Portuguese was developed based on Spanish and English resources. Also we have been working in the multilingual aspect, having built a translation system for entities from Arabic to English [207], which has led to a participation in the ACE07 competition. In the framework of the development of the Arabic WordNet Project [208][209] some efforts have been devoted to including Arabic NEs, and linking them to their corresponding English counterparts. A system for the assignation of alias to named entities applicable to information extraction systems have been built [198][199]. **Study and development of knowledge discovery tools and textual entailment.** In [10][55], new methods based on lexical and syntactical analysis were experimented. It has been applied to Answer Validation [44][45] and to automatic summarization [23][53]. Moreover, new techniques to semantic similarity determining based on dependency parsing trees were applied to textual entailment tasks [40][54]. In addition, the automatic extension of Spanish and English temporal reasoning systems was developed to other languages lacking resources like Catalan or Italian [29][30][34][35][36]. We have also been working in an automatic learning system (using Sams and AdaBoost) that uses distances based in semantics, with which UPC has participated in Pascal Challenge 07 [210]. Later on, this first system has been further enriched and has been used to participate in the 4th Recognizing Textual Entailment (RTE) Challenge Track at TAC 2008 (see [211]). Some experiments have been carried out in order to develop a complete system for automatic authorship identification based on Natural Language Processing techniques [14] [72]. **Adaptation of tools to specific domains.** A complete resource to textual representation based on logical forms has been developed. This mechanism allows the connection between domain-independent lexical resources and domain-dependent ones (like UMLS), guarantying the portability of tools to new specific domains [37]. Moreover, new models to semantic classifications have been defined to improve the robustness on semantic disambiguation systems, allowing their automatic adjustment to new domains [15][16][20]. In addition, we have integrated external knowledge from different resources (UMLS metathesaurus, CCHMC corpus, MeSH ontology…) in order to improve a multilabel text categorization system [113][105][108]. Furthermore, we have been working in geographical NERC (using gazeteers and ontologies), with participation in GEOCLEF 2007 (geographical search in a text corpus) and GEOQUERY 2007(question analysis and classification) [200] [201]. In 2008, TextMess research groups have jointly participated in the GEOCLEF 2008 contest by integrating the individual systems in the GeoTextMESS system using Fuzzy Borda for fusion of the individual results. This combination of systems is the result of a collaboration among the groups of the project participating in GEOCLEF (UJA, UPC, UPV) [34] [161]. Adaptation of methods for Named Entities Recognition in biomedical domain [143] [178].

Other useful tools. Development of AnCoraPipe, a tool for the annotation of corpora at all levels of analysis: morphology, syntax (constutuents and functions), Named Entities, WordNet synsets, thematic roles, semantic class, and correference [235]. Algorithm for the automatic identification of the correference chain first mentions [229], [241]. WaCOS: The Watermarking Corpus On-line System. WaCOS allows users the evaluation of corpus features (imbalance among categories, broadness domain, text length and writing style) http://www.dsic.upv.es/grupos/nle/demos.html. [195].

M2A3 - Multilingualism techniques.

This activity deals with new multilingual techniques (Spanish English Catalan) to be applied in the text mining tasks of the project. Techniques aimed at the minimization of errors that occurred during translation in multilingual text mining systems will be studied.

Results: Exploration of alternatives to machine translation based in lexical resources. Several alternatives based on multilingual semantic resources like the Interlingual Index module from EuroWordNet and Wikipedia have been tested in [43][67]. Comparison of two approaches to multilingual document clustering: one based on feature translation and another based on cognate identification [75]. **Evaluation, integration and refining machine translators.** A new Translation Module has been developed, that integrates several online machine translators and implements some heuristics to combine different translations [110][111]. A preliminary work based on Web frequencies to select the best translation for QA task has been developed [156]. **Development of mapping techniques between linguistic units from different languages.** These techniques were applied in temporal reasoning systems [31] [36]. In order to determine the similarity between two documents, we proposed a new approach based on a fuzzy system that tries to incorporate the human knowledge about the importance of named entities categories on documents [83].

M2A4 - Techniques to undertake interactivity.

In this activity the goal is to analyze study and experiment with interactivity techniques applied to text mining systems through two different means: through context and through dialog enrichment.

Results: Interactivity techniques through context. A method to automatic topic detection based on ML over a set of questions is being developed [40]. **Interactivity techniques through dialogue enrichment.** A new ML method to emotion classifying is used to detect the aim of the user, enriching the underlying information [48].

M2A5 - Techniques for processing non standard types of documents.

Currently, there is a huge variety of document formats. Some of them contain only raw text, while others contain marks to allow the identification of aspects of their structure (i.e. Websites, oral transcriptions and documents containing captions). TM technologies can be applied to any of these text types, whose processing involves the definition of a specific set of tasks: captions, oral transcriptions, and metadata.

Results: Techniques on captions. Several methods based on passage retrieval have been used to search information on captions [56]. Therefore, we have used techniques based in external knowledge [122][126][103][105] and information gain [95][100][104]. **Techniques on transcriptions.** A method to expand textual topics in a retrieval system using transcription of videos has been studied [123]. Moreover, information gain techniques has been applied to select the best transcriptions [121][102]. On the other hand, a method to classify video files making use of an information retrieval system [128]. Besides, advances in our work in these techniques are reflected firstly in the coordination, organization [202] and participation [203] in QAst 2007 competition. Later on, we have also co-ordinated, organized and participated in QAst 2008, producing

respectively publications [213] and [214]. On another area but also dealing with different types of documents, a work on summarisation has been carried out, producing a flexible multitask summarizer architecture that deals with documents in different languages, domains or media [223].

Module 3 - Intelligent, Interactive and multilingual Text Mining

This module consists on building the systems for Question Answering, Information Retrieval and Information Extraction. Annually, these systems will participate in the competitions in the stipulated deadlines for each one of them. Four different activities where planned to accomplish it.

M3A1 - Building the Question Answering systems "the need behind the query".

QA systems are the nearest paradigm to the natural way in which a human user expresses his/her information needs and receives the appropriate information. In this activity, our plan is to enhance a basic model of QA system (independent, factual-type, domain-independent, monolingual and textual questions) in several lines.

Results: Architecture of a QA system. A module to language-independent question classifying based on ML has been developed in [38]. A Textual Entailment Recognition module has been created and tested for the improvement of QA [11]. A multilingual passage retrieval for QA has been developed [7]. Finally, an Answer Extraction Module based on Semantic Roles has been created and incorporated in a general QA system (IBQA) [24][25][26][70]. A novel distributed multi-layer collaborative cache architecture for QA has been defined [215] [216]. **Monolingual and multilingual QA.** A Multilingual QA system based on EuroWordNet (ILI module) and Wikipedia has been obtained [43], and based on ontologies [12][13]. Also, AliQA has been adapted to multilingual texts [66]. Two approaches for multilingual question answering has been developed and compared, one based on merging passages and other merging answers [109] [182]. **Extension of the QA systems for their interactive use.** The efficacy of inference mechanisms based on syntactic information when applied to QA systems was tested in [32][65]. **Building of QA systems in restricted domains.** A QA system based on patterns has been applied to the healthcare domain [37]. Moreover, the adaptation of AliQA, an open-domain QA system, to the academic domain has been accomplished in [52].

M3A2 - Information Extraction.

The aim of Information Extraction (IE) is to identify and extract automatically the relevant information -according to a set of predefined rules - from a collection of texts that belong to a specific domain. In this activity, we will develop an IE system for the biomedical domain and the porting to different domains will be studied.

Results: Architecture of an Information Extraction system. A module for people search in web pages (Web People Search) has been developed, with a fine-grained person name categorization [21][22]. Furthermore a prototypical architecture for an IE system adaptable to new domains and based on perceptrons has been defined [217]. The system is able to extract mentions to relevant relationships among entities. A first prototype for the protein-protein interaction problem has been described in [158]. **Information Extraction system in the biomedical domain.** We have obtained a system to extract weak named entities and abbreviations in the biological domain [63].

M3A3 - Information Retrieval.

The aim of this activity is to define the architecture of the IR system, the HLT tools useful for the system and the necessary resources applied to it.

Results: Robust cross-language IR. We have integrated additional modules to the IR-n system (obtained in previous projects) to the improvement and adaptation to restricted domains.

Concretely, it was included a new indexing module optimized to the treatment of great volumes of information, and new modules to query expansion on image caption searching [28][56]. Also, IR-n has been expanded to work over a collection of Wikipedia images [57], and to deal with ambiguous words and word sense disambiguation in IR tasks [60]. **Study and implementation of techniques for combining the results in Image Retrieval systems.** Several experiment has been carry out in order to adapt IR-n System to multimodal texts [58] [71]. It has been adapted to multimodal corpora in medical domain [59]. Besides different techniques have been used in order to improve the final results combining text with images [122][99][100][126][102][103][104], videos [123][128] and transcriptions [121][102]. **Cross Language Spoken document Retrieval** [121][102]. The IR system on IBQA [70] has been adapted to deal with transcription texts (QAst task) [64]. **Multilingual web retrieval.** A self-training method for text categorization, authorship attribution and WSD based on the Web and language independent has been developed [141][142]. **Geographic Information Retrieval** [106][107][129][130].

M3A4 - Clustering, visualization, exploration and synthesis of search results.

In this activity, we will study how to cluster, display and facilitate to the user the exploration and synthesis of the searching results, taking ranked results, founded by one or several engines, as starting point.

Results: Organization and displaying of the results. A new system to web page clustering based on name disambiguation (UA-ZSA) has been developed in [49]. **Classification of results and documents.** Interesting results have been obtained by integrating linguistic information as features for text categorization tasks. The use of some information at higher level (Part of Speech, lemmatization and other combinations) has been found useful as enriched features in the automatic categorization of documents, aiming new studies in this direction [120]. Also, the LVQ algorithm successfully applied in multi-class text categorization has been tested as binary classifier under the Adaptive Selection of Base Classifiers approach [119] for multi-labeling text categorization, identifying suitability of certain algorithms to this kind of approach. A complete search engine featuring clustering and visualization of search results based on Formal Concept Analysis (FCA) has been developed as a result of a PhD dissertation [92] (demo at http://bender.lsi.uned.es:8080/ModuloWeb/jbraindead.html).

Module 4 - International Evaluation Campaigns

The goal of this module is to connect the project with the main international evaluation campaigns for search systems and HLT. On the one hand, we will participate in these campaigns with the aim of contrasting our results with those of the best international research groups; on the other, we will promote and coordinate some tasks with two goals: promoting research on the basic lines of this project and ensuring the presence, in this competitive field and being on an equal footing, of the languages of interest in this project, Catalan and Spanish. This module constitutes the core of the validation and evaluation activities of the project. Three activities were designed to achieve it.

M4A1 - Design of measures and evaluation methods.

This activity aims at the development of evaluation methods suitable for Text Mining tasks handled by the project.

Results: A new measure to the evaluation of open-domain QA systems with time restrictions was obtained. This measure was used in the real-time QA pilot task of the CLEF competition [27][62]. Our researches in non-parametric evaluation frameworks for combining metrics in information synthesis tasks have lead to develop general models for combining evaluation metrics without requiring weighting schemes (see [73]). In addition, we have developed a new model for estimating

the quality of a clustering system in the context of an interactive information retrieval system. An evaluation Framework in the area of definitional QA has been defined [218]. As coordinators of the QAst competition, new metrics for QA in Speech Transcripts have been defined [219].

M4A2 - Organization of international campaigns for competitive evaluation.
The goal in this activity is to take part in the main international evaluation campaigns, in order to ensure their compatibility with the tasks proposed within the project and to focus our research around them.
Results: Some of the participants in this project are members of the CLEF (Cross Language Evaluation Forum) Steering Committee (Julio Gonzalo – UNED, Anselmo Peñas – UNED, and José Luis Vicedo – UA) during 2007 and 2008 editions. UNED team has been involved in the organization of CLEF tracks since the beginning of the evaluation forum. In 2008 UNED has taken part in the coordination and organization of the following tracks: Answer Validation Exercise (AVE) [90]; Interactive Cross-Language Retrieval (iCLEF) [78], [81], [84]; Multiple Language Question Answering (QA@CLEF) [79]. In addition, UNED has organized other tracks in the framework of other competitive evaluation such as Web People Search Evaluation Workshop (WEPS-2) to be held in the context of the World Wide Web Conference (WWW2009). UPC group has organized QAst (Question Answering in Speech Transcripts) 2007 and 2008 competitions; has been in charge of the synthesis and evaluation of the results in task 9 of SemEval 2007 competitions, consisting in the annotation in several levels of Catalan and Spanish. The description of the task can be found in[220]; and has participated in the organization of the CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies (see [221][221]).

M4A3 - Participation in international evaluation campaigns.
In this activity we deal with direct evaluation of the systems built in the project, by means of their comparison with those of the most competitive laboratories at an international level.
Results: The following participations were accomplished: CLEF 2007 (Cross Language Evaluation Forum): Monolingual Ad-Hoc task [61] [182], Answer Validation (AVE) task [42], ImageCLEFphoto07 [56], ImageCLEFmedical07 [122], CL-SR07 [121], GeoCLEF [111] [181] [181] [200], GEOQuery [201], QAst [203].; SEMEVAL 2007 (Workshop on Semantic Evaluations): [179][180][222]. Web People Search task [49], Affective Text [48], Coarse-grained english all words [46], Multilevel Semantic Annotation of Catalan and Spanish [246], [237], [238], [239], [243]; ACL-PASCAL 2007 (Workshop On Textual Entailment And Paraphrasing): Textual Entailment Recognition (RTE) task [40] [220]; TREC 2007 (Text REtrieval Conference): GENOMICS task [63]; EVALITA 2007 (Evaluation of NLP Tools for Italian): Named Entity Recognition task [17], Temporal Expression Recognition and Normalization task [35] [29]; CLEF 2008 (Cross Language Evaluation Forum): Answer Validation (AVE) task [45], ImageCLEFphoto [58], Robust WSD task [60][162] , Medical Retrieval Task [4] [59], WikipediaMM task [57], Question Answer in Spoken Texts (QAST) [64] [214], iCLEF [85], WebCLEF [77], ImageCLEFmedical08 [126], VideoCLEF08 [128], GeoCLEF08 [129][130][160][161]; ACE 2007: entity mention detection (emd) and relationship mention detection (rmd) tasks [217]; entity translation task (specifically from Arabic to English) [207]; TAC's RTE (Recognizing Textual Entailment) 2008. [211]; Recognising Textual Entailment Challenge (RTE) [88].

Module 5 - Dissemination of results

Results were timely disseminated through international evaluation campaigns, and journal and conference contributions as shown at the end of this document. Moreover, the TEXT-MESS Wiki

http://gplsi.dlsi.ua.es/text-mess includes a public space where social agents and general public can receive information about the project achievements. In addition, to increase its social impact, the project essentials were promoted among social media press: the press agency EFE produced in April 2007 an interview with the Text-Mess coordinator (P. Martínez-Barco) that was released to several Spanish press media: *El Pais* (08/04/2007), *Información* (08/04/2007), *La Verdad* (08/04/2007), *ABC* (08/04/2007), *La Razón* (08/04/2007), as well as some electronic bulletins like *Fresqui.com* (07/04/2007), *Silicon News* (07/04/2007), *La Segunda Online* (07/04/2007), *Infobae Profesional.com* (09/04/2007), *MadrI+D* (09/04/2007). Also, a radio interview with Patricio Martínez-Barco was developed by Radio Nacional de España during *La Plaza* live radio program (17/04/2007). AnCora corpora and Lexicons were disseminated through different distribution lists (Linguist-List, ECOSEL, AESLA, Lingua-Net, TermLat) and web pages (RAE, UB).

3 Result indicators

Degree of achievement of the foreseen objectives: As pointed out in the last section, the first objective (G1) regarding the study and development of Text Mining systems (search, extraction, classification, information retrieval and text analysis) has been developed at 60%. The second one (G2), about improving and adapting already existing resources, techniques and tools, and creating new ones to TM systems, has been accomplished at 90%. And the last one (G3), promoting the research lines of the project in the main international evaluation campaigns, has been accomplished at 70%. Additionally, as there is still a year to go before the presentation of the final project report, the total attainment of the three objectives is viewed as an achievable aim. **Relevance and originality of the results**: Results have been qualitatively and quantitatively evaluated and periodically published. One of the main channels was participation in international evaluation competitions (achieving highly ranked positions) as well as publication in high-impact journals and conferences, both national and international, most of them appearing in the ISI-JCR listings. Also, we have participated in international competitions as providers of linguistic resources.
Development and management of the project: There have been no major deviations from the initial work plan. The UA team was the responsible for the coordination and management of the project. During these first two years, the coordination team was not only in charge of coordinating technical aspects, but also supervising the project planning within deadlines and ensuring suitable levels of communication between the various subprojects, by means of periodic meetings and other channels (see section 2, module 1). The other teams have contributed coordinating specific modules, activities and tasks according to the plan (see Figure 1). Moreover, the following teams were responsible of the organization of the meetings (UA, Alicante, Feb. 2007) (UB, Jun. 2007) (UPC, Oct. 2007) (UJA, May 2008), (UPV, Oct. 2008) and 1st TextMess Seminar (UPC, Oct. 2007).

Next, a detailed analysis of the results indicators related to each sub-project is shown.

3.1 UA Sub-Project

Scientific production: The UA team has produced 65 publications during the course of the project: 34 publications in journals or conference series with ISSN ([1], [7] to [39]), and 31 contributions to other conference proceedings ([2] to [5], [40] to [66]). **Usefulness of the results:** Some companies are interested in the project results: *Taller Digital S.A.* is interested in a competitive intelligence portal in order to trawl through the information related to technological businesses. *Oftex* is interested in text mining oriented to export trade. *Directive Soft* is a company

working on data mining, and they have a real interest in our text mining resources in order to apply them in their products. *Notarline S.A.* is interested in an information extraction system within the legal and notary domain. **Training of personnel**: Three PhD Thesis have been accomplished on the project topics [67][68][69] and three Master thesis [70][71][72]. Moreover, four students (Z. Kozareva, R. Muñoz-Terol, S.Vázquez and D. Tomás) are expected to get finished their PhD work during 2009. In addition, eleven students have joined the GPLSI group starting their PhD thesis, of which five are research assistants with State and Regional grants (FPI program), two research assistants from Latin America (on UA grants) and a research assistant from Eastern Europe (on a UA grant). **Benefits of the coordination**: In general, the coordination has enabled the comparison between different techniques applied to the same problem. Besides, the comparison results have been analyzed and corroborated at international competitions, in which some of the research groups of this project have participated. One of these competitions is CLEF, where UA has compared its techniques with some of the other teams at tracks: AVE 2007/2008 (UNED-UJA), ImageCLEFPhoto2007/2008 (UJA), Robust-WSD2008 (UJA), QAST2008 (UPC), ImageCLEFMed2008 (UJA). Besides, there are some joint publications showing collaborative efforts together with other Text-Mess teams [1][2][3][4][5]. In addition, the coordination has consolidated a collaborative work between the teams, building a robust framework of HLT resources, tools and systems applied to text mining called InTime [8]. **European and International collaborations**: The UA group is part of the FP6 QALL-ME Project consortium together with Foundazione Bruno Kessler (Italy, Prof. Bernardo Magnini), Univ. of Wolverhampton (UK, Prof. Ruslan Mitkov), DFKI (Germany, Prof. Gunter Newman), and the industrial partners (Comdata, S.p.A., Ubiest, S.p.A. and Waycom S.r.l.). This European project is developing Question Answering technologies in a multimodal and multilingual environment. Besides, our group has chaired international conferences, such as TIME 2007 and CLEF 2006, and we have participated in several committees of international conferences like CLEF 2007 and CLEF 2008. In addition, collaboration with other external international groups has been done, leading to several research stays: E. Saquete (Laboratory for Linguistics and Computation at Brandeis University in Massachusetts, USA, Prof. James Pustejovsky). Ó. Ferrández (Artificial Intelligence group at the International Computer Science Institute in Berkerley, USA, Prof. Srini Narayanan), Z. Kozareva (Information Sciences Institute in California, USA, Prof. Eduard Hovy).

3.2 UNED Sub-Project

Scientific production: The UNED team has produced 19 publications during the course of the project: 4 publications in journals or conference series with ISSN ([73] to [76]), and 15 contributions to other conference proceedings ([77] to [92]). **Usefulness of the results**: The Web People Search Task has attracted the attention of several IT companies: Spock (USA) has sponsored the workshop, and Google (USA) and Alias-i (USA) are participating in the Programme Committee for the task. The project has originated a research contract between UNED and Alma Technologies in the framework of an R&D project funded by the Ministerio de Industria. A workshop, coordinated by UNED, was held to bring together user communities and researchers in the field of Multilingual Information Access technologies. We received useful input from representatives belonging to patent offices, news agencies (JRC), government agencies and IT companies (Exalead and TextWise). **Training of personnel**: One PhD thesis relevant to the research lines of the project has been presented in 2008. Other two PhD work are expected to be finished in the following months. New researchers have joined the group, contributing with new approaches in the MLIA field. **Benefits of the coordination**: Text-Mess coordination has increased the awareness of Spanish research in the area within international research communities

and has increased the breadth and depth of Spanish research in the area. **European and International collaborations**: UNED has been involved in three EU-funded projects related to Text-Mess: Multimatch (Multilingual multimedia search engine in the Cultural Heritage domain), MedIEQ (Quality labelling of medical web content using multilingual information extraction) and TrebleCLEF (Evaluation, best practice and collaboration for Multilingual Information Access). In this year we have successfully finished MediEQ and Multimatch projects.

3.3 UJA Sub-Project

Scientific production: The UJA team has produced 40 publications during the course of the project: 23 publications in journals or conference series with ISSN ([93] to [115]), and 17 contributions to other conference proceedings ([116] to [131]). **Usefulness of the results**: Regarding a direct transference of the results obtained in this grant, and related issues, we have maintained contacts with several companies as *LYNKA*, *Novasoft*, *Natural Language*, which expressed its interest on the project proposal. We expect to implement the results to different applications in the future. Meanwhile we will maintain the flow of information and will attempt to have frequent contacts. **Training of personnel:** One PhD Thesis has been accomplished on the project topics [132]. Moreover, tree students (M. A. García-Cumbreras, M.C. Díaz and J.M. Perea) are expected to get finished their PhD work during 2009. In addition, one student has joined the SINAI group starting his PhD thesis, on an AECI program grant. **Benefits of the coordination**: The group has participated in all the general and specific meetings with members of all participating groups that took place during the first two years of the project. These meetings, on the one hand is used as a mechanism to test and control the development and management of the project, and secondly to discuss the modules and activities. The cross interaction of both groups has been very useful in these coordination meetings, since both groups have different backgrounds and skills. Coordinated participation in campaigns and international competitions, as GeoCLEF, ImageCLEF, ImageCLEFPhoto, VideoCLEF, QA&CLEF, Ad-Hoc, RTE, where we could compare the proposed systems with different approaches and techniques. **European and International collaborations:** Chaired and participation in several committees of international conferences. Collaboration with several groups: Ralf Steinberger, from European Commission Joint Research Centre -Institute for the Protection and Security of the Citizen (IPSC), Ispra, Italy (in fact the thesis of Arturo Montejo, member of group, was supervised by coordinator and Ralf Steinberger), or Manuel Montes and Luis Villaseñor from National Institute for Astrophysics, Optics and Electronics (INAOE- Mexico).

3.4 UPV Sub-Project

Scientific production: The UPV team has produced 62 publications during the course of the project: 26 publications in journals or conference series with ISSN ([133] to [158]), and 36 contributions to other conference proceedings ([159] to [194]). **Usefulness of the results**: The *Colegio Oficial de Médicos de Valencia*, as EPO, is interested in the results of the biomedical information extraction system. **Training of personnel:** One PhD Thesis has been conducted on one of the project topics [195]. Moreover, two students (Y. Benajiba and D. Buscaldi) are expected to finish their PhD work during 2009. R. Danger obtained a Juan de la Cierva post-doctoral grant (2008-2010) and was incorporated in the team of the project. **Benefits of the coordination**: The coordinated work has allowed us to share resources and develop a framework for the resources integration. Moreover, various systems have been integrated and we participated together with the team of the Jaén University at GeoCLEF 2008 [160]. **European and International Collaborations**: The results obtained in the project allowed the participation of some of the

members of the UPV group in several committees of international conferences. Various are the collaborations that the group has with others research institutes and universities in the context of PhD co-supervisions on: semi-supervised text categorization, with the INAOE in Puebla (Mexico) [141][142]; bio-inspired short text clustering with the *Universidad Nacional de San Luís* (Argentina) [135][140][157][168]; and, works about question answering in Arabic language with the *Ecole Mohammadia d'Ingenieurs* Rabat (Morocco) [172][173][174][175].

3.5 UPC Sub-Project

Scientific production: The UPC group has produced 28 publications since the project started. Out of these ones, 8 are publications in journals with ISSN ([196] to [203]), and 20 are publications in international conference proceedings ([204] to [223]). The exhaustive list can be found at the end of this document. Moreover, several systems (a summarizing system, a QA system for written and transcribed documents, an information extraction system and a geographical information retrieval system) have been developed, as described in section 2. **Usefulness of the results**: The company *Gestion del Conocimiento* is interested in the results of our research. **Training of personnel**: One PhD thesis related to the project topics has been presented, namely the one of M. Fuentes (2008). Furthermore, 5 students (E. González, P. R. Comas, D. Ferres, J. Poveda and E. Sapena) have registered their PhD thesis on topics directly related to the project and are being supervised by researchers from the project. Out of them, 3 are expected to finish their PhD thesis in 2009. **Benefits of the coordination**: As described in section 2, there have been several fruitful collaborations with the other groups of the project, such as the ones in paraphrases learning (UB), geographical IR (UPV, UJA) and oral QA (UPV). **European and International Collaborations:** Our international collaborations in the frame of the project include the collaboration with Princeton University in order to build the Arabic WordNet, as well as with both the *Laboratoire d'Informatique pour la Mécanique et les Sicences de l'Ingenieur* (LIMSI) and ELDA in order to organize QAst competitions. Several members of the group have been members of the program committee in several international conferences (H. Rodríguez, A. Ageno and J. Turmo in IJCAI 07 and 08, J. Turmo in CBA08). Furthermore, there have been the research stays of D. Ferrés at the ISI of University of Southern California (with the group of E. Hovy), E. González in New York University (with the group of Satoshi Sekine), and of J. Poveda in the University of Sheffield (with the group of Yorick Wilks).

3.6 UB Sub-Project

Scientific production: During the first two years of the project the UB team has produced 24 publications in journals or conference series with ISSN ([224] to [247]), and two PhD thesis [248][249]. **Usefulness of the results**: Several firms are interested in our resources: we have a contract with Microsoft (contract NVJ #NVJ1010110021), and we have participated with THERA in different R&D projects: Gamilen (FIT-350300-2006-93), Trujiman (FIT-330100-2006-198). **Training of personnel**: Two PhD Thesis have been accomplished on the project topics [248][249]. Moreover, three students, M. Recasens (FPU, AP2006-00994), A. Peris (FPU, AP2007-01028) and M. Vila (FI) are currently working on their PhD thesis inside the project topics. **Benefits of the coordination**: Being CLiC a group of linguists the participation in a coordinated group with computer scientist allows us to test and improve our linguistic resources and to receive the necessary technical support and computer knowledge. **European and International Collaborations:** The CLiC group is one of the organizers of the CoNLL-2009 Shared task Competition, SemEval 2010, and ARE-2009. We have organized the international CBA workshop on Corpus Based Approaches to Discourse Analysis, November 2008.

4 References

[1] Martínez-Barco, P., Palomar, M., Gonzalo, J., Peñas, A., Ureña, L.A., Martín, M.T., Pla, F., Rosso, P., Ageno, A., Turmo, J., Martí, M.A., Taulé, M. TEXT-MESS: Minería de Textos Inteligente, Interactiva y Multilingüe basada en Tecnología del Lenguaje Humano. Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), num. 41, pp. 317-318, ISSN 1135-5948, 2008, Spain.

[2] Gómez J. M., Rosso P., Sanchis E. Re-ranking of Yahoo snippets with the JIRS Passage Retrieval system. Proc. Workshop on Cross Lingual Information Access, CLIA-2007, 20th Int. Joint Conf. on Artificial Intelligence, IJCAI-07, Hyderabad, India, January 6-12, 2007.

[3] Gómez J. M., Buscaldi D., Rosso P., Sanchis E. JIRS Language-independent Passage Retrieval system: A comparative study. Proc. 5th Int. Conf. on Natural Language Processing, ICON-2007, Hyderabad, India, January 4-6, 2007.

[4] Navarro, S.; Díaz M.C.; Muñoz, R; García, M.A.; Llopis, F.; Martín M.T.; Ureña A.L.; Montejo A. "Text-mess in the Medical Retrieval ImageCLEF08 Task", CROSS LANGUAGE EVALUATION FORUM, Aarhus, September 2008.

[5] Navarro, S.; García, M.A.; Llopis, F.; Díaz M.C.; Muñoz, R; Martín M.T.; Ureña A.L.; Montejo A. "Text-mess in the ImageCLEFphoto08 Task", CROSS LANGUAGE EVALUATION FORUM, Aarhus, September 2008.

[6] Buscaldi, D., Perea, J.M., Rosso, P., Ureña, A., Ferrés, D., Rodríguez, H. GeoTextMESS: Result Fusion with Fuzzy Borda Ranking in Geographical Information Retrieval Geoclef 2008. November, 2008.

[7] Gómez, J.M. Recuperación de Pasajes Multilingüe para la Búsqueda de Respuestas. Sociedad Española para el Procesamiento del Lenguaje Natural, 40, pp. 149-152, 2008. ISSN 1135-5948

[8] Gómez, J.M. InTiMe: Plataforma de Integración de Recursos de PLN. Procesamiento del Lenguaje Natural, 40, pp. 83-90, ISSN 1135-5948, 2008.

[9] Ferrández, Ó., Micol, D., Muñoz, R., Palomar, M. Técnicas léxico-sintácticas para el reconocimiento de Implicación Textual. Procesamiento del Lenguaje Natural, nº 38. ISSN: 1135-5948. pp. 53-60. April, 2007.

[10] Ferrández, Ó., Micol, D., Muñoz, R., Palomar, M. DLSITE-1: Lexical Analysis for Solving Textual Entailment Recognition. NLDB '07, Lecture Notes in Computer Science, 4592, ISSN: 0302-9743. pp. 284-294 June, 2007.

[11] Ferrández, O., Muñoz, R., Palomar, M., Improving Question Answering Task by Textual Entailment Recognition. Lecture Notes in Computer Science, NLDB 2008, num. 5039, pp. 339-340, 2008.

[12] Ferrández, O., Izquierdo, R., Ferrández, S., Vicedo, JL. Addressing ontology-based question answering with collection of user queries. Information Processing and Management (In press).

[13] Ferrández, O., Izquierdo, R., Ferrández, S., Vicedo, JL. Un sistema de Búsqueda de Respuestas basado en ontologías, implicación textual y entornos reales. Procesamiento del Lenguaje Natural, 41, pp. 47-54. 2008. ISSN 1135-5948.

[14] Guillén–Nieto, V.; Vargas–Sierra, Ch.; Pardiño–Juan, M., Martínez–Barco, P.; Suárez–Cueto, A. Exploring state-of-the-art software for forensic authorship identification. IJES, International journal of english studies, vol. 8, num. 1, pp. 1-28, 2008, Spain. ISSN 1578-7044.

[15] Izquierdo R., Suárez A. and Rigau G. A Proposal of Automatic Selection of Coarse-grained Semantic Classes for WSD. Proceedings of the 23th Annual Meeting of Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN'07. Sevilla, España. Procesamiento del Lenguaje Natural num. 39 pp. 189-196. ISSN: 1135-5948. 2007.

[16] Izquierdo, R., Suárez, A., Rigau, G. Exploring the Automatic Selection of Basic Level Concepts. In Proceedings of Recent Advances in NLP (RANLP07).

[17] Kozareva Z.; Montoyo A. UNIALI: Building a resource independent system for Italian named entity recognition. Intelligenza artificiale, vol 4, nº 2, pp. 73-74, June, 2007.

[18] Kozareva Z.; Vázquez S.; Montoyo A. Multilingual Name Disambiguation with Semantic Information. Lecture Notes in Computer Science 4629, pp. 23-30. 2007

[19] Kozareva, Z.; Vázquez, S.; Montoyo A. The influence of context during the categorization and discrimination of Spanish and Portuguese person names. Procesamiento del Lenguaje Natural, 39, pp. 81-88. 2007

[20] Kozareva, Z.; Vázquez S.; Montoyo A. The Usefulness of Conceptual Representation for the Identification of Semantic Variability Expressions. Lecture Notes in Computer Science 4394, 2007.

[21] Kozareva, Z., Moraliyski, R., Dias, G. Web People Search with Domain Ranking. Lecture Notes in Computer Science. ISBN:978-3-540-87390-7, TSD 2008, Czech Republic.

[22] Kozareva, Z., Vázquez, S., Montoyo, A. Domain Information for Fine-Grained Person Name Categorization. Lecture Notes in Computer Science. ISSN: 0302-9743. pp: 311-321. Cicling 2008, Israel.

[23] Lloret, E., Ferrández, O., Muñoz, R., Palomar, M. Integración del reconocimiento de la implicación textual en tareas de automáticas de resúmenes de textos. Procesamiento del Lenguaje Natural, 41, pp. 183-190, ISSN 1135-5948, 2008.

[24] Moreda, P., Llorens, H., Saquete, E., Palomar, M. Automatic generalization of a QA Answer Extraction Module based on Semantic Roles. Lecture Notes in Computer Science. IBERAMIA 2008, pp. 233-242. ISSN 0302-9743.

[25] Moreda, P., Llorens, H., Saquete, E., Palomar, M. Two proposals of a QA Answer Extraction Module based on Semantic Roles. Lecture Notes in Computer Science. MICAI 2008, pp. 174 -184.

[26] Moreda, P., Llorens, H., Saquete, E., Palomar, M. The influence of Semantic Roles in QA: a comparative análisis. Procesamiento del Lenguaje Natural, 41, pp. 55-62. ISSN 1135-5948.
[27] Noguera, E.; Llopis. F; Ferrández A.; Escapa A. New Measures for Open-Domain Question Answering Evaluation Within a Time Constraint. Proc. 10th Int. Conf. on Text, Speech and Dialogue, TSD-2007, Springer-Verlag, LNAI (4629), pp. 540-547, 2007.
[28] Noguera, E.; Llopis, F. Passage Retrieval vs. Document Retrieval in the CLEF 2006 Ad Hoc Monolingual Tasks with the IR-n System. Revised Selected Papers CLEF-2006, Springer-Verlag, LNCS(4730), 2007.
[29] Puchol-Blasco, M., Saquete, E., Martínez-Barco, P. Il sistema RETA2 per l'italiano. Intelligenza artificiale, vol 4, nº 2, pp. 60-61, June, 2007.
[30] Puchol-Blasco, M., Saquete, E., Martínez-Barco, P. Aprendizaje automático para el reconocimiento temporal multilingüe basado en TiMBL. Proceedings of the 23th Annual Meeting of Sociedad Española para el Procesamiento del Lenguaje Natural, SEPLN'07. Sevilla, España. Procesamiento del Lenguaje Natural num. 39 pp. 97-104. ISSN: 1135-5948. 2007.
[31] Puchol-Blasco, M., Saquete, E., Martínez-Barco, P. Multilingual Extension of Temporal Expression Recognition using Parallel Corpora. 14th International Symposium of Temporal Representation and Reasoning. IEEE Computer Society, P2836, pp. 175-180, June, 2007.
[32] Roger, S., Ferrández, A., Peral, J., Ferrández, S., López-Moreno, P. An Inference Mechanism for Question Answering. Journal of Computer Science & Technology (JCS&T). ISSN: 1666-6038. pp. 21-27. March, 2007.
[33] Romá-Ferri, M.T.; Hermida, J.M., Montoyo, A., Palomar, M. Representación del conocimiento farmacoterpéutico: diseño de una ontología. XI Congreso Nacional de Informática de la Saluda (INFORSALUD 2008). Madrid, Sociedad Española de Informática de la Salud (SEIS), 15-17 de abril de 2008: 240-247.
[34] Saquete, E., Martínez-Barco, P.,Muñoz, R. Evaluation of an Automatic Extension of Temporal Expression Treatment to Catalan. Lecture Notes in Computer Science, 4394, pp. 166-174. 2007.
[35] Saquete, E., Martínez-Barco, P., Muñoz Guillena, R. Il sistema TERSEO per l'italiano. Intelligenza artificiale, vol 4, nº 2, pp. 62-63, June, 2007
[36] Saquete, E, Ferrández, O., Ferrández, S., Martínez-Barco, P., Muños, R., Combining automatic acquisition of knowledge with machine learning approaches for multilingual temporal recognition and normalization. Information Sciences, num. 17, vol. 178, pp. 3319-3332, ISSN 0020-0255, 2008.
[37] Terol, R.M., Martínez-Barco, P., Palomar, M. A knowledge based method for the medical question answering problem. Computers in Biology and Medicine. Vol. 37, Num. 10, pp. 1511-1521. October, 2007.
[38] Tomás, D.; Vicedo, José L. Multiple-Taxonomy Question Classification for Category Search on Faceted Information. Proc. 10th Int. Conf. on Text, Speech and Dialogue, TSD-2007, Springer-Verlag, LNAI (4629), pp. 653-660, 2007.
[39] Tomás, D.; Vicedo, José L.; Bisbal, E.; Moreno, L. TrainQA: a Training Corpus for Corpus-Based Question Answering Systems. Proc. 8th Int. Conf. on Computational Linguistics and Intelligent Text Processing, CICLing-2007, IEEE Computer Society, 2007.
[40] Boldrini, E., Martínez-Barco, P., Navarro, B., Puchol-Blasco, M., and Vargas, C. AQA: A multilingual anaphora annotation model for question answering. CBA 2008 Corpus-Based Approaches to Coreference Resolution in Romance Languages. Barcelona, 2008.
[41] Ferrández, Ó., Micol, D., Muñoz, R., Palomar, M. A Perspective-based Approach for Solving Textual Entailment Recognition. ACL PASCAL Workshop on Textual Entailment and Paraphrasing, pp. 66-71, June, 2007.
[42] Ferrández, Ó., Micol, D., Muñoz, R., Palomar, M. The contribution of the University of Alicante to AVE 2007. In on-line Working Notes, CLEF 2007, Budapest, Hungary, September, 2007.
[43] Ferrández, S.; Ferrández, O.; Ferrández, A.; Muñoz, R. The Importance of Named Entities in Cross-Lingual Question Answering. Int. Conf. Recent Advances in Natural Language Processing, RANLP-2007, September 27-29, 2007.
[44] Ferrández, O., Muñoz, R., Palomar, M. TE4AV: Textual Entailment for Answer Validation. 2008 IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'08). October 19 – 22, 2008 Beijing, China.
[45] Ferrández, O., Muñoz, R., Palomar, M. A Lexical-Semantic Approach to AVE. Cross Language Evaluation Forum (CLEF 2008), Aarhus, Septiembre 2008.
[46] Izquierdo, R., A. Suarez and G. Rigau: GPLSI: Word coarse-grained Disambiguation aided by Basic Level Concepts. Proceedings of the fourth international workshop on semantic evaluations (SemEval 2007). Association for Computational Linguistics, 157-160. Prague, June 2007
[47] Kozareva, Z., Vazquez, S. and Montoyo, A. Discovering the Underlying Meanings and Categories of a Name through Semantic and Domain Information. International Conference Recent advance in Natural Language Processing 2007.
[48] Kozareva, Z., Navarro, B., Vazquez, S., and Montoyo, A. UA-ZBSA: A Headline Emotion Classification through Web Information. International Workshop on Semantic Evaluations SEMEVAL. 2007
[49] Kozareva Z.; Vázquez S.; Montoyo A.. UA-ZSA: Web Page Clustering on the basis of Name Disambiguation. International Workshop on Semantic Evaluations SEMEVAL. 2007

[50] Kozareva, Z.; Vázquez, S.; Montoyo, A. A Language Independent Approach for Name Categorization and Discrimination, Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, Association for Computational Linguistics, Prague, Czech Republic, , pp. 19--26., 2007.

[51] Kozareva, Z.; Riloff, E. and Hovy, E. Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs. in Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, (ACL-08: HLT), Columbus, USA, June 2008.

[52] López-Moreno, P.; Ferrández, A.; Roger, S.; Ferrández, S. The problems in a Question Answering system in the academic domain. Int. Conf. Recent Advances in Natural Language Processing, RANLP-2007, pp. 395-399, September 27-29, 2007.

[53] Lloret, E., Ferrández, O., Muñoz, R., Palomar, M. A Text Summarization Approach Under the Influence of Textual Entailment. 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2008), pages 22–31, 12-16 June, Barcelona, Spain.

[54] Micol, D., Ferrández, Ó., Muñoz, R., Palomar, M. DLSITE-2: Semantic Similarity Based on Syntactic Dependency Trees Applied to Textual Entailment. HLT-NAACL Workshop TextGraphs-2: Graph-Based Methods for Natural Language Processing, pp. 73-80, April, 2007.

[55] Micol, D., Ferrández, Ó., Muñoz, R., Palomar, M. A Semantic-less Approach for the Textual Entailment Recognition Task. Int. Conf. Recent Advances in Natural Language Processing, RANLP-2007, September 27-29, 2007.

[56] Navarro, S.; Llopis, F.; Muñoz, R.; Noguera, E. Information Retrieval of Visual Descriptions with IR-n System based on Passages. In on-line Working Notes, CLEF 2007, Budapest, Hungary, September, 2007.

[57] Navarro, S.; Muñoz, R; Llopis, F. "A Textual Approach Based on Passages Using IR-n in WikipediaMM Task 2008", Cross Language Evaluation Forum (CLEF 2008), Aarhus, 2008.

[58] Navarro, S.; Llopis, F.; Muñoz, R. "Different Multimodal Approaches using IR-n in ImageCLEFphoto 2008", Cross Language Evaluation Forum (CLEF 2008), Aarhus, 2008.

[59] Navarro, S.; Muñoz, R; Llopis, F. "A Multimodal Approach to the Medical Retrieval Task using IR-n", Cross Language Evaluation Forum (CLEF 2008), Aarhus, 2008.

[60] Navarro, S.; Llopis, F.; Muñoz, R. "IRn in the CLEF Robust WSD Task 2008", Cross Language Evaluation Forum (CLEF 2008), Aarhus, 2008.

[61] Noguera, E.; Llopis, F. Applying Query Expansion Techniques to Ad Hoc Monolingual tasks with the IR-n system. 8th Int. Cross-Language Evaluation Forum CLEF-2007 Working Notes, Budapest, Hungary, September 19-21, 2007.

[62] Noguera, E.; Llopis. F; Ferrández A.; Escapa A. Exploring New Measures for Open-Domain Question Answering Evaluation within a Time Constraint. Int. Conf. Recent Advances in Natural Language Processing, RANLP-2007, September 27-29, 2007.

[63] Pardiño, M.; M. Terol R.; Martínez-Barco P.; Llopis F.; Noguera E. Using IR-n for Information retrieval of Genomics Track. TREC 2007, November 6-9, 2007.

[64] Pardiño, M.; Gómez, J.M.; Llorens, H.; Muñoz-Terol, R.; Navarro, B.; Saquete, E.; Martínez-Barco, P.; Moreda, P.; Palomar, M. "Adapting IBQAS to work with Text Transciptions in QAst Task: IBQAst", Cross-Language Evaluation Forum (CLEF 2008), Aarhus, Septiembre 2008.

[65] Roger, S.; Ferrández, A.; Peral, J.; Ferrández, S; López-Moreno, P. Un mecanismo de inferencia aplicado a la búsqueda de respuesta. XII Congreso Argentino de Ciencias de la Computación, 2006.

[66] Muñoz-Terol, R., Puchol-Blasco, M., Pardiño, M., Gómez, J.M., Roger, S., Vila, K., Ferrández, A., Peral, J. and Martínez-Barco, P. AliQAn, Spanish QA System at Multilingual QA@CLEF-2008 Cross-Language Evaluation Forum (CLEF 2008), Aarhus, Septiembre 2008.

[67] Ferrández Escámez, Sergio. Arquitectura multilingüe de sistemas de búsqueda de respuestas basada en ILI y Wikipedia. Supervised by Antonio Ferrández. Universidad de Alicante, June 2008.

[68] Moreda Pozo, Paloma. Los roles semánticos en la tecnología del lenguaje humano: Anotación y Aplicación. Supervised by Manuel Palomar. Universidad de Alicante, July 2008.

[69] Navarro Colorado, Fco. de Borja. Metodología, construcción y explotación de corpus anotados semántica y anafóricamente. Phd thesis. Supervised by Manuel Palomar and Patricio Martínez-Barco. Universidad de Alicante, September 2007.

[70] Llorens, Héctor. Búsqueda de respuestas basada en conocimiento semántico. Master thesis. Supervised by Estela Saquete. Universidad de Alicante, December 2008.

[71] Navarro, Sergio. M^2 IR-n: Sistema de RI Multimodal y Multidominio basado en Pasajes. Master thesis. Supervised by Fernando Llopis and Rafael Muñoz. Universidad de Alicante, December 2008.

[72] Pardiño, María. Aplicación de técnicas de procesamiento de lenguaje natural a la atribución de autoría. Master thesis. Supervised by Armando Suárez. Universidad de Alicante, December 2008.

[73] Amigó, E., Gonzalo, J., Artiles, J. & Verdejo, F. "A comparison of extrinsic clustering evaluation metrics based on formal constraints". Information Retrieval Journal, 2008

[74] López-Ostenero, F., Peinado, V., Gonzalo, J. & Verdejo, F. "Interactive Question Answering: Is Cross-Language Harder than Monolingual Searching?". Information Processing and Management. Special topic issue on User-centered Evaluation of Information Retrieval Systems, 2008, Vol. 44, pp. 66-81

[75] S. Montalvo, R. Martínez, A. Casillas, V. Fresno. "Multilingual news clustering: Feature translation vs. identification of cognate named entities". Pattern Recognition Letters. Volumen: 28. Pág.: 2305 -- 2311. ISSN: 0167-8655, 2007. Elsevier B.V.

[76] Peñas, A.; Rodrigo, Á.; Sama, V. & Verdejo, F. Testing the Reasoning for Question Answering Validation. Journal of Logic and Computation. Oxford University Press 2008. Volumen: 18(3) Special Issue: Natural Language and Knowledge Representation 459-474. Online ISSN 1465-363X. Print ISSN 0955-792X

[77] Amigó, E.; Martínez-Romo, J.; Araujo, L. & Peinado, V. "UNED at WebCLEF 2008: Applying High Restrictive Summarization, Low Restrictive Information Retrieval and Multilingual Techniques". Evaluating Systems for Multilingual and Multimodal Information Access. 9th Workshop of the Cross-Language Evaluation Forum (CLEF 2008). LNCS. Springer Verlag.

[78] Clough, P., Gonzalo, J., Karlgren, J., Barker, E., Artiles, J. & Peinado, V. "Large-Scale Interactive Evaluation of Multilingual Information Access Systems - the iCLEF Flickr Challenge". Workshop on Novel Methodologies for Evaluation in Information Retrieval. 30th European Conference on Information Retrieval (ECIR 2008). 2008

[79] P. Forner, A. Peñas, I. Alegria, C. Forăscu, N. Moreau, P. Osenova, P. Prokopidis, P. Rocha, B. Sacaleanu, R. Sutcliffe, E. Tjong, K.Sang. "Overview of the CLEF 2008 Multilingual Question Answering Track". Evaluating Systems for Multilingual and Multimodal Information Access. 9th Workshop of the Cross-Language Evaluation Forum (CLEF 2008). LNCS. Springer Verlag.

[80] Giampiccolo, D.; Forner, P.; Peñas, A.; Ayache, C.; Jijkoun, V.; Osenova, P.; Rocha, P.; Sacaleanu, B. & Sutcliffe, R. In Peters, C.; Jijkoun, V.; Mandl, T.; Müller, H.; Oard, D.; Peñas, A.; Petras, V. & Santos, D. *(ed.)* Overview of the CLEF 2007 Multilingual Question Answering Track. Advances in Multilingual and Multimodal Information Retrieval, CLEF 2007, Revised Selected Papers. Lecture Notes in Computer Science. Springer-Verlag, 2008, Vol. 5152, ISSN: 0302-9743

[81] J. Gonzalo, P. Clough, J. Karlgren. "Overview of iCLEF 2008: Search Log Analysis for Multilingual Image Retrieval". Evaluating Systems for Multilingual and Multimodal Information Access. 9th Workshop of the Cross-Language Evaluation Forum (CLEF 2008). LNCS. Springer Verlag.

[82] Marlow, J., Clough, P., Cigarrán Recuero, J. and Artiles, J. (2008), Exploring the Effects of Language Skills on Multilingual Web Search, In Proceedings of the 30th European Conference on IR Research (ECIR'08), Glasgow, UK, April 2008, LNCS4956, pp. 126-137.

[83] S. Montalvo, R. Martínez, A. Casillas, V. Fresno. "Bilingual News Clustering Using Named Entities and Fuzzy Similarity". Text, Speech and Dialogue. Lecture Notes in Artificial Intelligence, subseries of Lecture Notes in Computer Science (LNCS), ISSN 0302-9743, ISBN 978-3-540-74627-0. Volumen: 4628. Pág.: 107—114, 2007.

[84] Peinado, V., Artiles, J., Gonzalo, J., Barker, E. & López-Ostenero, F. "FlickLing: a Multilingual Search Interface for Flickr". Working Notes for the CLEF 2008 Workshop. 2008

[85] Peinado, V., Gonzalo, J., Artiles, J. & López-Ostenero, F. "UNED at iCLEF 2008: Analysis of a large log of multilingual image searches in Flickr". Evaluating Systems for Multilingual and Multimodal Information Access. 9th Workshop of the Cross-Language Evaluation Forum (CLEF 2008). LNCS. Springer Verlag.

[86] Peñas, A.; Rodrigo, Á. & Verdejo, F. In Peters, C.; Jijkoun, V.; Mandl, T.; Müller, H.; Oard, D.; Peñas, A.; Petras, V. & Santos, D. (ed.) Overview of the Answer Validation Exercise 2007 Advances in Multilingual and Multimodal Information Retrieval, CLEF 2007, Revised Selected Papers. Lecture Notes in Computer Science. Springer-Verlag, 2008 Volumen: 5152 ISSN: 0302-9743

[87] A. P. Garcia-Plaza, V. Fresno, R. Martínez. "Web Page Clustering Using a Fuzzy Logic Based Representation and Self-Organizing Maps". Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence, Pág.: 851 - 854. ISBN: 978-0-7695-3496-1. DOI 10.1109/WIIAT.2008.249.

[88] Rodrigo, Á., Peñas, A. & Verdejo, F. "Towards an Entity-based recognition of Textual Entailment". Text Analysis Conference (TAC) 2008 Workshop. Maryland, USA. 2008.

[89] Téllez-Valero, A., Montes-y-Gómez, M., Villaseñor-Pineda, L. & Peñas, A. "Improving Question Answering by Combining Multiple Systems via Answer Validation". Gelbukh, A. F. (ed.). Computational Linguistics and Intelligent Text Processing (CICLing 2008). Springer-Verlag, 2008

[90] Rodrigo, Á.; Peñas, A. & Verdejo, F. In Peters, C.; Jijkoun, V.; Mandl, T.; Müller, H.; Oard, D.; Peñas, A.; Petras, V. & Santos, D. (ed.) UNED at Answer Validation Exercise 2007. Advances in Multilingual and Multimodal Information Retrieval, CLEF 2007, Revised Selected Papers. Lecture Notes in Computer Science. Springer-Verlag. 2008 Volumen: 5152 ISSN: 0302-9743.

[91] Rodrigo, Á., Peñas, A., Verdejo, F. "Overview of the Answer Validation Exercise 2008". Evaluating Systems for Multilingual and Multimodal Information Access. 9th Workshop of the Cross-Language Evaluation Forum (CLEF 2008). LNCS. Springer Verlag.

[92] Cigarrán, J. M. Organización de Resultados de Búsqueda mediante Análisis Formal de Conceptos. Tesis Doctoral. Universidad Nacional de Educación a Distancia. 2008.
[93] Martín, M.T., Díaz, M.C., Montejo, A., Ureña, L.A. Integración de conocimiento en un dominio específico para categorización multietiqueta. Procesamiento del Lenguaje Natural (ISSN: 1135-5948). 63-70.
[94] García-Cumbreras, M.A., Martín, M.T., Ureña, L.A., Díaz, M.C., Montejo, A. Using translation heuristics to improve a multimodal and multilingual retrieval information system. Application of Fuzzy Sets Theory. Vo. 4578 Lecture Notes in Computer Science.
[95] Díaz, M.C., Martín, M.T., Montejo, A., Ureña L.A.. Mejora de los sistemas multimodales mediante el uso de ganancia de información. 119-130. 2007. Procesamiento del Lenguaje Natural, 38. 119- 130. Abril 2007.
[96] Martín, M.T.; Ureña, L.A.; García-Vega, M. The learning vector quantization algorithm applied to automatic text classification tasks. Neural Network . Volumen: 20. 748-756. Ed Elsevier. 2007
[97] Martínez-Santiago, F., García-Cumbreras, M.A., Montejo, A. SINAI at CLEF 2006 Ad Hoc Robust Multilingual Track: query expansion using the Google search engine. LNCS Springer-Verlag. Volumen 4730. 2007.
[98] Martínez-Santiago, F., Montejo, A., García-Cumbreras, M.A. Representación formal de la estructura lógica de sitios web, y su aplicación a un navegador web multilingüe basado en diálogo. SEPLN, ISSN 1135-5948 nº38, Abril 2007.
[99] García-Cumbreras, M.A., Martín, M.T, Ureña L.A., Díaz, M.C., Montejo, A. Using translation heuristics to improve a multimodal and multilingual retrieval information system. Application of Fuzzy Sets Theory. Vol 4578. LNCS Springer-Verlag. 2007.
[100] Díaz, M.C.; García-Cumbreras, M.A., Martín, M.T., Montejo, A. Ureña, L.A. Using Information Gain to Improve the ImageCLEF 2006 Collection. Lecture Notes in Computer Science (LNCS). ISSN: 0302-9743. Vol 4730, pp 711-714. Springer-Verlag. 2007
[101] Martínez-Santiago, F. El virus Eliza. Ciencia Cognitiva. 2008
[102] Díaz, M.C., Martín, M.T., García-Cumbreras, M.A., Ureña, L.A. Using Information Gain to Filter Information in CLEF CL-SR Track. LNCS Springer-Verlag. ISSN: 0302-9743. 2008
[103] Díaz, M.C., García-Cumbreras, M.A., Martín, M.T., Montejo, A., Ureña, L.A. Integrating MeSH Ontology to Improve Medical Information Retrieval. LNCS Springer-Verlag. ISSN:0302-9743. 2008.
[104] Martín, M.T.; Díaz, M.C.; Montejo, A; Ureña, L.A. Using Information Gain to Improve Multimodal Information Retrieval Systems. Information Processing & Management. Vol 44 ,pp. 1146 - 1158. Elsevier. 2008.
[105] Martín, M.T.; Montejo, A., Díaz, M.C., Perea, J.M., Ureña, L.A. Expanding Terms with Medical Ontologies to Improve a Multi-Label Text Categorization System. En "Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration, editado por Prince and Roche. IGI Global publication 2009.
[106] Perea, J.M, García-Cumbreras, M.A., García-Vega, M., Ureña, L.A., Comparing Several Textual Information Retrieval Systems for the Geographical Information Retrieval task. 13th International Conference on Applications of Natural Language to Information Systems (NLDB 2008). LNCS Springer-Verlag. ISSN: 0302-9743. 2008.
[107] Perea, J.M, García-Cumbreras, M.A., García-Vega, M., Ureña, L.A. Sistemas de Recuperación de Información Geográfica multilingües en CLEF. Procesamiento del Lenguaje Natural, 40 . 129- 136. Abril 2008.
[108] Perea, J.M, Martín, MT., Montejo, A., Díaz, M.C. Categorización de textos biomédicos usando UMLS. Procesamiento del Lenguaje Natural, 40 . 121- 127. Abril 2008.
[109] Aceves-Pérez R.M.; Montes y Gómez M.; Villaseñor, L.; Ureña, L.A. Two Approaches for Multilingual Question Answering: Merging Passages vs. Merging Answers. International Journal of Computational Linguistics and Chinese Language Processing Special Issue on Cross-Lingual Information Retrieval and Question Answering . Vol. 13 N. 1, 2008.
[110] García-Cumbreras, M.Á., Díaz, M.C. Martín, M.T., Montejo, A., Ureña, L.A. SINAI System: Combining IR Systems at ImageCLEFPhoto 2007. Revised Selected Papers CLEF-2007, Springer-Verlag, LNCS(4730), 2007.
[111] Perea, J.M., García-Cumbreras, M.Á., García-Vega, M., Ureña, L.A. Filtering for improving the geographic information search. Revised Selected Papers CLEF-2007, Springer-Verlag, LNCS(4730), 2007.
[112] García-Cumbreras, M.Á., José M. Perea-Ortega, F. Martínez-Santiago, L.Alfonso Ureña-López. Combining Lexical Information with Machine Learning for Answer Validation at QA@CLEF 2007. Revised Selected Papers CLEF-2007, Springer-Verlag, LNCS(4730), 2007.
[113] A. Montejo, Ureña, L. A., García-Cumbreras, M.A., Perea, J. M.. Using Linguistic Information as Features for Text Categorization. Mining Massive Data Sets for Security pp. 245-254. F. Fogelman-Soulié et al (Eds.). ISBN: 978-1-58603-898-4. IOS Press (2008).
[114] Díaz, M.C.; Martín, M.T.; Ureña, L.A. Query Expansion with a Medical Ontology to Improve a Multimodal Information Retrieval System. Computers in Biology and Medicine Vol. (In Press) 44, 2009.
[115] Chapter III: Expanding Terms with Medical Ontologies to Improve a Multi-Label Text Categorization System. Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration Edited By: Violaine Prince and Mathieu Roche. Medical Information Science Reference 2009. ISBN-10: 1605662747

[116] García-Cumbreras M.A., Perea, J.M., Martínez-Santiago F., Ureña, L.A. "Combining Lexical Information with Machine Learning for Answer Validation at QA@CLEF 2007". Proceedings of CLEF (Cross Language Evaluation Forum) 2007.
[117] García-Cumbreras M.A., Díaz M.C., Martín, M.T., Montejo,A., Ureña, L.A."SINAI System: Combining IR Systems at ImageCLEFPhoto 2007" Proceedings of CLEF (Cross Language Evaluation Forum) 2007.
[118] Montejo, A., Martín, M.T, Ureña, L.A. Experiences with the LVQ algorithm in multilabel text categorization. Proceedings of the 4th International Workshop on Natural Language Processing and Cognitive Science, 12-13 June, 2007 - Funchal, Madeira - Portugal. ISBN: 978-972-8865-97-9, pp. 213-221.
[119] Montejo, A., Perea, J.M., Martinez-Santiago, F., García-Cumbreras, M.A., Martín, M.T., Ureña, L.A. Combining Lexical-Syntactic Information with Machine Learning for Recognizing Textual Entailment. Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Praga, Rep. Checa. 28-29 de Junio de 2007, pp. 78-82.
[120] Montejo, A., Ureña, L.A., García-Cumbreras, M.A., Perea, J.M.. Using Linguistic Information as Features for Text Categorization. Proceedings of the NATO Advanced. Study Institute on Mining Massive Data Sets for Security. Pages 107-108. 2007.
[121] Díaz, M.C., Martín, M.T., García-Cumbreras, M.A., Ureña, L.A. SINAI at CL-SR task at CLEF 2007. Proceedings of CLEF (Cross Language Evaluation Forum) 2007.
[122] Díaz, M.C., García-Cumbreras, M.A., Martín, M.T., Montejo, A., Ureña, L.A. SINAI AT IMAGECLEF 2007. Proceedings of CLEF (Cross Language Evaluation Forum) 2007.
[123] Díaz, M.C., Perea, J.M., Martín, M.T., Montejo, A., Ureña, L.A. SINAI at TRECVID 2007. TREC Video Retrieval Evaluation (TRECVID) 2007.
[124] Martinez-Santiago, F., García-Cumbreras, M.A., Montejo, A.. Applying Google search engine for robust cross-lingual retrieval. Proceedings of CLEF (Cross Language Evaluation Forum) 2007.
[125] Perea, J.M., García-Cumbreras, M.A., García-Vega, M., Montejo, A.. GeoUJA System. University Of Jaén At Geoclef 2007. Proceedings of CLEF (Cross Language Evaluation Forum) 2007.
[126] Díaz, M.C., García-Cumbreras, M.A., Martín, M.T., Ureña, L.A., Montejo, A. SINAI at IMAGECLEF 2008. Proceedings of CLEF (Cross Language Evaluation Forum) 2008.
[127] Martínez-Santiago, F., Perea, J.M., García-Cumbreras, M.A.. SINAI at Robust WSD Task @ CLEF 2008: When WSD is a good idea for Information Retrieval tasks?. Proceedings of CLEF (Cross Language Evaluation Forum) 2008.
[128] Perea, J.M., Montejo, A., Martín, M.T., Díaz, M.C., Ureña, L.A. SINAI at VideoCLEF 2008. Proceedings of CLEF (Cross Language Evaluation Forum) 2008.
[129] Perea, J.M., García-Cumbreras, M.A., García-Vega, M., Ureña, L.A. SINAI-GIR System. University of Jaén at GeoCLEF 2008. Proceedings of CLEF (Cross Language Evaluation Forum) 2008.
[130] Perea, J.M., Ureña, L.A., Buscaldi, D., Rosso, P. TextMESS at GeoCLEF 2008: Result Merging with Fuzzy Borda Ranking. Proceedings of CLEF (Cross Language Evaluation Forum) 2008.
[131] Montejo, A., Perea, J.M., Martinez-Santiago, F., García-Cumbreras, M.A., Martín, M.T., Ureña, L.A.; Combining Lexical-Syntactic Information with Machine Learning for Recognizing Textual Entailment. Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Praga, Rep. Checa. 28-29 de Junio de 2007, pp. 78-82.
[132] García Vega, M. Resolución de la ambigüedad léxica mediante aprendizaje por cuantificación vectorial, Universidad de Jaén. December 2006.
[133] Buscaldi D., Rosso P. A conceptual density-based approach for the disambiguation of toponyms. International Journal of Geographical Information Science, 22 (3): 143-153. ISSN: 1365-8816, 2008.
[134] Ferretti E., Errecalde M., Rosso P. Does Semantic Information Help in the Text Categorisation Task? Journal of Intelligent Systems, ISSN: 0334-1860, vol. 17, No. 1-3, pp.91-107, 2008.
[135] Ingaramo. D., Pinto D., Rosso P., Errecalde M. Evaluation of internal validity measures in short-text corpora. Proc. 9th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2008, Springer-Verlag, LNCS(4919), pp. 555-567, 2008.
[136] Guzmán R., Montes M., Rosso P., Villaseñor L. A Web-based Self-training Approach for Authorship Attribution. Proc. 6th Int. Conf. on Natural Language Processing, GoTAL-2008, Advances in Natural Language Processing, Springer-Verlag, LNCS(5221), pp. 160-168, 2008.
[137] Buscaldi D., Rosso P. On the Relative Importance of Toponyms in GeoCLEF, Revised Selected Papers CLEF-2007, Springer-Verlag, LNCS(5152), pp. 815-822, 2008.
[138] Buscaldi D., Benajiba Y., Rosso P. Sanchis E. Web-based Anaphora Resolution for the QUASAR Question Answering System, Revised Selected Papers CLEF-2007, Springer-Verlag, LNCS(5152), pp. 324-327, 2008.
[139] Danger R., Rosso, P., Pla, F., Molina, A. PPIEs: Protein-Protein Interaction Information Extraction system. Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), num. 40, pp. 137-143, 2008.
[140] Ingaramo. D., Errecalde M., Rosso P. Density-based clustering of short-text corpora. Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), num. 41, 2008.

[141] Guzmán R., Montes M., Rosso P., Villaseñor L. Improving Text Classification by Web Corpora. Proc. 5th Atlantic Web Intelligence Conf., AWIC-2007, Advances in Intelligent Web Mastering, Advances in Software Computing, vol. 43, Springer-Verlag, pp. 154-159, 2007.
[142] Guzmán R., Montes M., Rosso P., Villaseñor L. Taking advantage of the Web for Text Classification with imbalanced classes. In: Proc. 6th Mexican International Conference on Artificial Intelligence, MICAI-2007, Springer-Verlag, LNAI (4827), pp. 831-838, 2007.
[143] Ponomareva N., Pla F., Molina A., Rosso P. Biomedical Named Entity Recognition: A poor knowledge HMM-based approach. Proc. 12th Int. Conf. on Applications of Natural Language to Information Systems, NLDB-2007, Springer-Verlag, LNCS(4593), pp. 382-387, 2007.
[144] Benajiba Y., Rosso P., Benedí J.M. ANERsys: An Arabic Named Entity Recognition system based on Maximum Entropy. Proc. 8th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2007, Springer-Verlag, LNCS(4394), pp. 143-153, 2007.
[145] Benajiba Y., Rosso P., Gómez J.M. Adapting JIRS Passage Retrieval System to the Arabic. Proc. 8th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2007, Springer-Verlag, LNCS(4394), pp. 530-541, 2007.
[146] Pinto D., Benedí J.M., Rosso P. Clustering Narrow-Domain Short Texts by using the Kullback-Leibler distance. Proc. 8th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2007, Springer-Verlag, LNCS(4394), pp. 611-622, 2007.
[147] Buscaldi D., Rosso P. Some experiments in Humour Recognition using the Italian Wikiquote collection. Proc. 3rd Workshop on Cross Language Information Processing, CLIP-2007, Int. Conf. WILF-2007, Springer-Verlag, LNAI(4578), pp. 464-468, 2007.
[148] Buscaldi D., Rosso P., Sanchis E. A WordNet-based indexing technique for Geographical Information Retrieval. Evaluation of Multilingual and Multi-modal Information Retrieval, Revised Selected Papers CLEF-2006, Springer-Verlag, LNCS(4730), 2007.
[149] Buscaldi D., Gómez J. M., Rosso P., Sanchis E. N-gram vs. Keyword-based Passage Retrieval for Question Answering. Evaluation of Multilingual and Multi-modal Information Retrieval, Revised Selected Papers CLEF-2006, Springer-Verlag, LNCS(4730), 2007.
[150] Buscaldi D., Rosso P. A Bag-of-words based Ranking method for the Wikipedia Question Answering task, Revised Selected Papers CLEF-2006, Springer-Verlag, LNCS(4730), 2007.
[151] Pinto D., Rosso P., Jiménez E. A Penalisation-Based Ranking approach for the mixed monolingual task of WebCLEF 2006, Revised Selected Papers CLEF-2006, Springer-Verlag, LNCS(4730), pp. 826-829, 2007.
[152] Pinto D., Rosso P. On the Relative Hardness of Clustering Corpora, Proc. 10th Int. Conf. on Text, Speech and Dialogue, TSD-2007, Springer-Verlag, LNAI (4629), pp. 155-161, 2007.
[153] Pinto D., Juan A., Rosso P. Using Query-Relevant Documents Pairs for Cross-Lingual Information Retrieval, Proc. 10th Int. Conf. on Text, Speech and Dialogue, TSD-2007, Springer-Verlag, LNAI (4629), pp. 630-637, 2007.
[154] Alexandrov. M., Blanco X., Ponomareva N., Rosso P. Constructing empirical models for automatic dialogue parametrization, Proc. 10th Int. Conf. on Text, Speech and Dialogue, TSD-2007, Springer-Verlag, LNAI (4629), pp. 455-463, 2007.
[155] Levner E., Pinto D., Rosso P., Alcaide D., Sharma R. Fuzzifying clustering algorithms: The case study of MajorClust, Proc. 6th Mexican International Conference on Artificial Intelligence, MICAI-2007, Springer-Verlag, LNAI, 2007.
[156] Rosso. P., Buscaldi D., Iskra M. Web-based selection of optimal translations of short queries, Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), num. 38, pp.49-52, 2007.
[157] Ingaramo. D., Errecalde M., Rosso P. Medidas internas y externas en el agrupamiento de resúmenes científicos de dominios reducidos, Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), num. 39, pp. 55-62, 2007.
[158] Danger, Roxana; Rosso, Paolo; Ferran, Pla, Molina, Antonio. PPIEs: Protein-Protein Interaction Information Extraction system. Revista SEPLN No. 40
[159] Benajiba Y., Diab M., Rosso P. Arabic Named Entity Recognition using Optimized Feature Sets. Proc. Int. Conf. on Empirical Methods in Natural Language Processing, EMNLP-2008, Waikiki, Honolulu, U.S.A., October 2008
[160] Perea J.M., Ureña L.A., Buscaldi D., Rosso P. TextMESS at GeoCLEF 2008: Result Merging with Fuzzy Borda Ranking. 9th Int. Cross-Language Evaluation Forum CLEF-2008 working notes, Aarhus, Denmark, September 2008.
[161] Buscaldi D., Rosso P. The UPV at GeoCLEF 2008: The GeoWorSE System. 9th Int. Cross-Language Evaluation Forum CLEF-2008 working notes, Aarhus, Denmark, September 2008.
[162] Buscaldi D., Rosso P. QA with a Disambiguated Document Collection. 9th Int. Cross-Language Evaluation Forum CLEF-2008 working notes, Aarhus, Denmark, September 2008.
[163] Pinto D., Civera J., Juan A., Rosso P., Barrón A. A statistical approach to crosslingual natural language tasks. Proc. 4th Latin American Workshop on Non-Monotonic Reasoning, LANMR-2008, Puebla, Mexico, October 2008.
[164] Cagnina L., Errecalde M., Ingaramo D., Rosso P. A discrete particle Swarm optimizer for clustering short-text corpora. Proc. Bioinspired Optimization Methods and their Applications, BIOMA-2008, Ljubljana, Slovenia, October 2008.

[165] Buscaldi D., Rosso P. Map-based vs. Knowledge-based Toponym Disambiguation. Proc. 5th Int. Workshop on Geographical Information Retrieval, GIR-2008, CIKM-2008, Napa Valley, U.S.A., October 2008.
[166] Buscaldi D., Rosso P. Geo-WordNet: Automatic Georeferencing of WordNet. Proc. 6th Int. Conf. on Language Resources and Evaluation, LREC-2008, Marrakech, Morocco, May 2008.
[167] Benajiba Y., Rosso P. Arabic Named Entity Recognition using Conditional Random Fields. Proc. Workshop on HLT & NLP within the Arabic world. Arabic Language and local languages processing: Status Updates and Prospects, 6th Int. Conf. on Language Resources and Evaluation, LREC-2008, Marrakech, Morocco, May 2008.
[168] Errecalde M., Ingaramo D., Rosso P. Proximity estimation and the hardness of short-text corpora. 5th Workshop on Text-based Information Retrieval, TIR-2008, In: Proc. Database and Expert Systems Applications, DEXA-2008, IEEE Press, Turin, Italy, September 2008.
[169] Barrón A., Rosso P., Pinto D., Juan A. On Cross-lingual Plagiarism Analysis using a statistical model. 2nd Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse, PAN-2008, In: Proc. 18th European Conf. on Artificial Intelligence, ECAI-2008, Patras, Greece, July 2008.
[170] Barrón A., Rosso P. Towards the exploitation of statistical Language Models for Plagiarism detection with reference. 2nd Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse, PAN-2008, In: Proc. 18th European Conf. on Artificial Intelligence, ECAI-2008, Patras, Greece, July 2008.
[171] Benajiba Y., Diab M., Rosso P. Arabic Named Entity Recognition: An SVM-based approach. Proc. Int. Arab Conf. on Information Technology, ACIT-2008, Hammamet, Tunisia, 2008
[172] Abouenour L., Bouzoubaa K., Rosso P. Improving Q/A using Arabic WordNet. Proc. Int. Arab Conf. on Information Technology, ACIT-2008, Hammamet, Tunisia, 2008
[173] Abouenour L., Bouzoubaa K., Rosso P. Towards an Arabic Q/A system using a conceptual/lexical ontology (in Arabic). Proc. 5th Conf. on Scientific Research Outlook & Technology Development in the Arab world, SROV, Fez, Marroco, October 2008.
[174] Abouenour L., Bouzoubaa K., Rosso P. Construction de l'ontologie Amine Arabic WordNet dans le cadre des systèmes Q/A (in French). Proc. 2nd JOurnées Scientifiques en Technologies de l'Information et de la Communication JOSTIC-2008, Rabat, Marroco, 2008.
[175] Abouenour L., Bouzoubaa K., Rosso P. Système de Question/Réponse dans le cadre d'une plateforme intégrée: cas de l'Arabe (in French). Proc. Rencontre Nationale en Informatique : Outils et Applications, RINOA-2008, Errachidia, Morocco, June 2008.
[176] Roshchina A., Cardiff J., Rosso P., Troussov A. Ontology data freshness on the Social Web. Proc. 3rd Int. Conf. for Internet Technology and Security Transactions, ICITST-2008, Dublin, Ireland, June 2008.
[177] Perez F., Pinto D., Rosso P., Cardiff J. Constructing ontologies for narrow domains: Methodology for term extraction and relationship discovery. Proc. 3rd Int. Conf. for Internet Technology and Security Transactions, ICITST-2008, Dublin, Ireland, June 2008
[178] Ponomareva N., Rosso P., Pla F., Molina A. Conditional Random Fields vs. Hidden Markov Models in a biomedical Named Entity Recognition task. Int. Conf. Recent Advances in Natural Language Processing, RANLP-2007, pp. 479-483, September 27-29, 2007.
[179] Buscaldi D., Rosso P. UPV-WSD : Combining different WSD Methods by means of Fuzzy Borda Voting. Proc. Int. SemEval Workshop, Association of Computational Linguistics, Prague, Czech Republic, pp. 434-437, June 23-24, 2007.
[180] Pinto D., Rosso P., Jiménez H. UPV-SI: Word Sense Induction using Self-Term Expansion. Proc. Int. SemEval Workshop, Association of Computational Linguistics, Prague, Czech Republic, pp. 430-433, June 23-24, 2007.
[181] Buscaldi D., Rosso P. The UPV at GeoCLEF 2007. 8th Int. Cross-Language Evaluation Forum CLEF-2007 working notes, Budapest, Hungary, September 19-21, 2007.
[182] Buscaldi D., Benajiba Y., Rosso P., Sanchis E. The UPV at QA@CLEF 2007. 8th Int. Cross-Language Evaluation Forum CLEF-2007 working notes, Budapest, Hungary, September 19-21, 2007.
[183] Buscaldi D., Rosso P. A Comparison of methods for the automatic identification of locations in Wikipedia. Proc. 4th Int. Workshop on Geographical Information Retrieval, GIR-2007, ACM CIKM-2007, Lisbon, Portugal, November 9, 2007.
[184] Benajiba Y., Rosso P., Lyhyaoui A. Implementation of the ArabiQA Question Answering System's components. Proc. Workshop on Arabic Natural Language Processing, 2nd Information Communication Technologies Int. Symposium, ICTIS-2007, Fez, Morroco, April 3-5, 2007.
[185] Benajiba Y., Rosso P. Towards a measure for Arabic corpora quality. Proc. Int. Colloquium on Arabic Language Processing, CITALA-2007, Rabat, Morroco, June 18-19, 2007.
[186] Benajiba Y., Rosso P. ANERsys 2.0 : Conquering the NER task for the Arabic language by combining the Maximum Entropy with POS-tag information. Proc. Workshop on Natural Language-Independent Engineering, 3rd Indian Int. Conf. on Artificial Intelligence, IICAI-2007, Pune, India, December 17-19, 2007.

[187] Pinto D., Rosso P., Benajiba Y., Ahachad A. Word Sense Induction in the Arabic Language: A Self-Term Expansion Based Approach. Proc. 7th Conf. on Language Engineering, The Egyptian Society Of Language Engineering, ESOLE-2007, Cairo, Egypt, December 5-6, 2007.

[188] Gómez J. M., Rosso P., Sanchis E. Re-ranking of Yahoo snippets with the JIRS Passage Retrieval system. Proc. Workshop on Cross Lingual Information Access, CLIA-2007, 20th Int. Joint Conf. on Artificial Intelligence, IJCAI-07, Hyderabad, India, January 6-12, 2007.

[189] Gómez J. M., Buscaldi D., Rosso P., Sanchis E. JIRS Language-independent Passage Retrieval system: A comparative study. Proc. 5th Int. Conf. on Natural Language Processing, ICON-2007, Hyderabad, India, January 4-6, 2007.

[190] Mascardi V., Rosso P., Cordì V. Enhancing communication inside multi-agent systems: An approach based on alignment via Upper Ontologies. Proc. Int. Conf. on Agents, Web-Services, and Ontologies: Integrated Methodologies, AWESOME007, Durham, UK, September 6-7, 2007.

[191] Mascardi V., Cordì V., Rosso P. A comparison of Upper Ontologies. Proc. Conf. on Agenti e industria: Applicazioni tecnologiche degli agenti software, WOA07, Genova, Italy, September 24-25, 2007.

[192] Mesmoudi M., De Floriani L., Rosso P. Theoretical foundations of 3D scalar field visualization. Proc. Int. Conf. VISAPP-2007, Barcelona, Spain, pp. 69-77, March 8-11, 2007.

[193] Rosso P., Benajiba Y. Towards an Arabic Question Answering system (in Arabic). Proc. 4th Conf. on Scientific Research Outlook & Technology Development in the Arab world, SROIV, Damascus, Syria, 11-14 December, 2006.

[194] Díaz, I., Moreno, L.. Metamorfosis: Aplicando técnicas de procesamiento de lenguaje natural para la deducción automática de interacciones. 33 Conferencia Latinoamericana de estudios de informática (CLEI 2007) ISBN 978-9968-9678-9-1. San José de Costa Rica (Costa Rica), 2007

[195] Pinto, D. On Clustering and Evaluation of Narrow Domain Short-Text Corpora. PhD. UPV.2008. Supervised by Paolo Rosso

[196] M. Surdeanu, Ll. Márquez, X. Carreras, P.R. Comas. Combination Strategies for Semantic Role Labeling. Journal of Artificial Intelligence Research 29, ISSN: 1076-9757, 2007.

[197] E. González, J. Turmo. Comparing Non-parametric Ensemble Methods for Document Clustering. Natural Language and Information Systems, pg 245—256, ISSN: 0302-9743. June, 2008.

[198] E. Sapena, Ll. Padró, J. Turmo. Alias Assignment in Information Extraction. Procesamiento del Lenguaje Natural, num. 39 pp. 105-112. ISSN: 1135-5948. 2007.

[199] E. Sapena, Ll. Padró, J. Turmo. A Graph Partitioning Approach to Entity Disambiguation Using Uncertain Information. Advances in Natural Language Processing, pg. 428—439, ISSN: 0302-9743, 2008.

[200] D. Ferrés, H. Rodríguez. TALP at GeoCLEF 2007: Results of a Geographical Knowledge Filtering Using Terrier. Advances in Multilingual and Multimodal Information Retrieval. Lecture Notes in Computer Science, vol. 5152, pg.830—833, ISSN: 0302-9743. September, 2008.

[201] D. Ferrés, H. Rodríguez. TALP at GeoQuery 2007: Linguistic and Geographical Analysis for Query Parsing. Advances in Multilingual and Multimodal Information Retrieval. Lecture Notes in Computer Science, vol. 5152, pg.834-833, ISSN: 0302-9743. September, 2008.

[202] J. Turmo, P.R. Comas, C. Ayache, D. Mostefa, S. Rosset, L. Lamel. Overview of QAST 2007. . Advances in Multilingual and Multimodal Information Retrieval. Lecture Notes in Computer Science, vol. 5152, pg. 249—256, ISSN: 0302-9743. September, 2008.

[203] P. R. Comas, J. Turmo, M. Surdeanu. Robust Question Answering for Speech Transcripts Using Minimal Syntactic Analysis. Advances in Multilingual and Multimodal Information Retrieval. Lecture Notes in Computer Science, vol. 5152, pg. 424—432, ISSN: 0302-9743. September, 2008.

[204] M. Surdeanu, R. Morante, Ll. Marquez. Analysis of Joint Inference Strategies for the Semantic Role Labeling of Spanish and Catalan. Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), February 2008.

[205] N. Català, M. Martin. Feature Selection for Support Vector Machines by Alignment with Ideal Kernel. Research Report LSI-07-48-R.

[206] J. Poveda, M. Surdeanu, J. Turmo. A Comparison of Statistical and Rule-Induction Learners for Automatic Tagging of Time Expressions in English. Proceedings of the 14th International Symposium on Temporal Representation and Reasoning, Alicante, June 2007.

[207] D. Farwell, J. Gimenez, E. González, R. Halkoum, H. Rodríguez, M. Surdeanu. The UPC System for Arabic-to-English Entity Translation. Proceedings of the NIST 2007 Automatic Content Extraction Workshop (ACE07), March 2007.

[208] H. Rodríguez, D. Farwell, J. Farreres, M. Bertran, M. Alkhalifa, Mª A. Martí, W.J. Black, S. Elkateb, J. Kirk, A. Pease, P. Vossen, C. Fellbaum. Arabic WordNet: Current State and Future Extensions. Proceedings of the Fourth International GlobalWordNet Conference - GWC 208, Szeged, Hungary. January, 2008.

[209] H. Rodríguez, D. Farwell, J. Farreres, M. Bertran, M. Alkhalifa, Mª A. Martí. Arabic WordNet: Semi-automatic Extensions using Bayesian Inference. Proceedings of the the 6th Conference on Language Resources and Evaluation LREC2008. Marrakech (Morocco). May 2008.

[210] D. Ferrés, H. Rodríguez. Machine Learning with Semantic-Based Distances between Sentences for Textual Entailment. ACL-PASCAL Workshop on Textual Entailment and Paraphrasing RTE 2007. Prague, June 2007.
[211] A. Ageno, F. Cruz, D. Farwell, D. Ferrés, H. Rodríguez, J. Turmo. TALP at TAC 2008: A semantic approach to Recognizing Textual Entailment. TAC 2008 Working Notes. November, 2008.
[212] P. R. Comas, J. Turmo. Spoken Document Retrieval Based on Approximated Sequence Alignment. Text, Speech and Dialogue, 11th International Conference, TSD 2008. September 200.
[213] J. Turmo, P.R. Comas, S. Rosset, L. Lamel, N. Moreau, D. Mostefa,. Overview of QAST 2008. 9th International Cross-Language Evaluation Forum CLEF-2008 Working Notes. Aarhus, Denmark, September 2008.
[214] P. R. Comas, J. Turmo. Robust Question Answering for Speech Transcripts: UPC Experience in QAst 2008. Working notes of Cross Language Evaluation Forum (CLEF). September, 2008.
[215] D. Domínguez-Sal, J.Ll. Larriba-Pey, M. Surdeanu. A Multi-layer Collaborative Cache for Question Answering. Proceedings of Euro-Par 2007, August 2007.
[216] D. Dominguez-Sal, M. Surdeanu, J. Aguilar-Saborit, J. Ll. Larriba-Pey. Cache-aware Load Balancing for Question Answering. Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM), 2008.
[217] M. Surdeanu, M. Ciraramita. Robust Information Extraction with Perceptrons. Proceedings of the NIST 2007 Automatic Content Extraction Workshop (ACE07), March 2007.
[218] S. Kanaan, J. Turmo. An Evaluation Framework based on Gold Standard Models for Definition Question Answering. Proceedings of the 5th International Conference on Natural Language Processing. Hyderabad, India, 2007.
[219] L. Lamel, S. Rosset, C. Ayache, D. Mostefa, J. Turmo, P. R. Comas. Question Answering on Speech Transcripts: The QAST evaluation in CLEF. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC), 2008.
[220] Ll. Màrquez, L. Villarejo, M.A. Martí, M. Taulé. SemEval-2007 Task 9: Multilevel Semantic Annotation of Catalan and Spanish. Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), June 2007.
[221] M. Surdeanu, R. Johansson, A. Meyers, Ll. Marquez, J. Nivre. The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies. Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL-2008), 2008.
[222] Ll. Márquez, Ll. Padró, M. Surdeanu, L. Villarejo. UPC: Experiments with Joint Learning within SemEval Task 9. SemEval 2007, June 2007.
[223] M. Fuentes. A Flexible Multitask Summarizer for Documents from Different Media, Domain, and Language. PhD. Thesis, Universitat Politècnica de Catalunya. March, 2008.
[224] Civit, M., M.A. Martí, N. Bufi (2006) 'Cat3LB and Cast3LB: from Constituents to dependencies'. Advances in Natural Language Processing (LNAI, 4139), pp. 141-153. Springer Verlag, Berlin. ISSN: 0302-9743.
[225] Martí, M.A., M. Taulé, Ll. Màrquez y M. Bertran (2007) 'Anotación semiautomática con Papeles Temáticos de los corpus CESS-ECE', Procesamiento del Lenguaje Natural-TIMM, Alicante.
[226] Martí, M. A. y M. Taulé (2007) 'CESS-ECE: corpus anotados del español y catalán', en Arena Romanistica. A new Nordic journal of Romance studies, núm. 1 Monografía dedicada a Corpus and text linguistics in Romance languages.
[227] Martí, M.A., M. Taulé, M. Arévalo (2006) 'MICE: un mòdul per al reconeixement i classificació d'entitats amb nom', a V. Salvador i L. Climent (eds). El discurs prefabricat II, Edicions Universitat Jaume I.
[228] Taulé, M.; Castellví, J.; Martí, M.A.; Aparicio, J. (2006) 'Fundamentos teóricos y metodológicos para el etiquetado semántico de CESS-CAT y CESS-ESP'. Procesamiento del Lenguaje Natural, Vol. 36. Alacant, Espanya. ISSN: 1135-5948.
[229] Recasens, M., M.A. Martí y M. Taulé. (2008). 'First-mention Definites: More than Exceptional Cases'. S. Featherston i S. Winkler (eds.), Linguistic Evidence: Empirical, Theoretical and Computational Perspectives. Berlin: De Gruyter. (En prensa).
[230] Recasens, M., M. Antònia Martí y M. Taulé (2007) 'Text as Scene: Discourse Deixis and Bridging Relations'. Procesamiento del Lenguaje Natural, n. 39. ISSN:1135-5948.
[231] Vallbé, J., M.Antònia Martí, Blaz Fortuna, Aleks Jaulin, Dunja Mladenic, Pompeu Casanovas (2007) 'Stemming and Lemmatization: Improving Knowledge Management through Language Processing Techniques'. Trends in Legal Knowledge, the Semantic Web and the Regulation of Electronic Social Systems.
[232] Morante, R. (2008) 'Etiquetat automàtic de rols semàntics amb un sistema basat en memòria'. Digithum, vol. 10.
[233] Aparicio, J., M. Taulé y M.A. Martí (2008) 'AnCora-Verb: A Lexical Resource for the Semantic Annotation of Corpora', Language Resources and Evaluation Conference (LREC'08), Marrakesh, Marroc.
[234] Aparicio, J., M. Taulé, M. y M.A. Martí (2008) 'AnCora-Verb: Two large-scale verbal lexicons for Catalan and Spanish', Proceedings of the 13th EURALEX International Congress. Barcelona, Espanya.
[235] Bertran, M., O. Borrega, M. Recasens, B. Soriano. (2008) 'AnCoraPipe: A tool for multilevel annotation'. Procesamiento del Lenguaje Natural.
[236] Borrega, O., M. A. Martí y M. Taulé (2007) 'What do we mean when we talk about Named Entities?', Corpus Linguistics, Birmingham. UK.

[237] Morante, R. (2008) 'Semantic role labeling tools trained on the Cast3LB-CoNLL-SemRol corpus'. /Language Resources and Evaluation Conference/ (LREC'08), Marrakesh, Marroc.
[238] Morante, R., B. Busser (2007) 'ILK2: Semantic Role Labelling for Catalan and Spanish using TiMBL', Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pp. 183-186, Prague, Czec Republic. Association for Computational Linguistics, ACL.
[239] Morante, R., A. van den Bosch (2007) 'Memory-based semantic role labeling'. Proceedings of the International Conference Recent Advances in Natural Language Processing, pp. 388-394, Borovets, Bulgary.
[240] Recasens, M., M. A. Martí, M. Taulé (2007) 'Where Anaphora and Coreference Meet. Annotation in the CESS-ECE Corpus'. Recent Advances in Natural language Processing. Borovets (Bulgaria).
[241] Recasens, M. (2009) 'A Corpus-driven Chain-starting Classifier of Definite NPs in Spanish', EACL, Athens, Greece
[242] Rodríguez, H., D. Farwell, J. Farreres, M. Bertran, Musa Alkhalifa, M. A. Martí, W. Black, S. Elkateb, J. Kirk, A. Pease, P. Vossen & Ch. Fellbaum (2008) 'Arabic WordNet: Current State and Future Extensions', WordNet Global Conference 2008. Szeged, Hungary.
[243] Surdeanu, M., R. Morante, y Ll. Màrquez (2008). 'Analysis of Joint Inference Strategies for the Semantic Role Labeling of Spanish and Catalan'. Computational Linguistics and Intelligent Text Processing 9th International Conference, CICLing-2008, pp. 206-218 Haifa, Israel, February.
[244] Taulé, M., M.A. Martí y M. Recasens (2008) 'AnCora: Multilevel Annotated Corpora for Catalan and Spanish', Language Resources and Evaluation Conference (LREC'08), Marrakesh, Marroc.
[245] Taulé, M.; Martí M.A; Castellví, J. (2006) 'Semantic Classes in CESS-LEX: Semantic Annotation of CESS-ECE'. Treebanks and Linguistics Theories, TLT-2006. Praga, Txecoslovàquia. ISBN: 80-239-8009-2.
[246] Màrquez, Ll. L. Villarejo, M. A. Martí and M. Taulé (2007) 'SemEval-2007 Task 09: Multilevel Semantic Annotation of Catalan and Spanish', Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 42–47, Prague. Association for Computational Linguistics, ACL.
[247] Martí, M.A., Duran, J., Perea, P. (2007) 'HistoCat y DialCat: extensiones de un analizador morfológico para tratar textos históricos y dialectales del catalán'. Procesamiento del Lenguaje Natural, n. 39, Sevilla (Spain).
[248] Morante, R. 'Computing meaning in interaction.' PhD Dissertation, Tilburg University.
[249] Szmidt Sierykow, D.T. 'La sonoritat en els grups consonàntics polonesos'. PhD Dissertation, U. Barcelona 2006.