Jornadas de Seguimiento de Proyectos, 2007
Programa Nacional de Tecnologías Informáticas

TIEMPO: IT Handling of Multilingual Financial and Economic Information TIN2004-07246-C03

M. Vilares Ferro [*] Univ. of Vigo
M.A. Alonso Pardo [†] Univ. of A Coruña
V.J. Díaz Madrigal [‡] Univ. of Sevilla

Abstract

The phenomenon of globalisation in the access to information forces us to face the challenge of adapting to it. There are four main reasons for this. Firstly, as a result of the overwhelming avalanche of relevant information now available. Secondly, due to the often critical nature of the decision-making process. Thirdly, as a consequence of a specific work environment in terms of the terminology, syntax and semantics employed. Finally, caused by the need to manage information in various languages. In this context, financial markets are one of the most sensitive sectors.
Keywords: Information recovery, knowledge acquisition, named entity recognition, natural language processing, multi-lingual access, parsing, tagging, term extraction, text alignment.

1 Objectives and project aim

Our aim is the development of an environment for extracting, managing and evaluating information from texts concerning financial markets in multi-lingual environments. As a complement, we contemplate a multi-lingual search for information structured around a natural language interface, and the generation of linguistic resources.

Such a proposal translates into a complex task requiring the conjunction of efforts in a number of specialized domains with particular goals that we can classify in three different levels: lexical, syntactic, and semantic. More in detail, our strategy includes a lexical phase whose aim is to identify named entities, terms and collocations in the text. Here, we have considered both available free tools and others developed by the research groups themselves.

The syntactic phase follows the lexical one, and its goal is to identify relevant components and relations between them, in the text. In order to deal with ungrammatical inputs, commonly found in the field of *natural language processing* (NLP), we consider shallow and robust parsing techniques. Resulting partial trees will be the basis for the semantical phase, in particular in order to allow automatic knowledge acquisition.

[*]Email: vilares@uvigo.es
[†]Email: alonso@udc.es
[‡]Email: vjdiaz@lsi.us.es

The multi-lingual character of our proposal leads us to provide strategies to pass efficiently from one language to another in the domain of *information retrieval* (IR) facilities. Having discarded the use of complex machine translation tools, we are exploring other alternatives, such as text alignment, which we could consider not only in the treatment of multi-lingual IR activities on parallel texts, but also as a basis for generating glossaries and dictionaries of interest to index generation in this context. Finally, the lack of multi-lingual corpora in the languages and domain in question, in particular in dealing with parallel texts, has obliged us to generate these linguistic resources ourselves.

2 Working tasks and success level

2.1 Lexical level

Tagging Tools. Tagging provides most of the information needed to identify and classify tokens from documents, which implies disambiguating lexical forms. We have used a variety of taggers in our tests. MrTagoo, an Hmm-based tagger, currently available for Spanish and Galician, has proved to be an adequate experimental platform to deal with IR related tools [1, 26, 49, 50], additionally including new operational capabilities such as spelling correction [17, 16, 20, 22, 21] and synonymy management facilities [45]. In order to develop generic strategies, we are also using TreeTagger and TnT, both free tools.

Automatic corpus annotation. The availability of large tagged data corpora is an essential aspect in most NLP tasks. The effort required to manually tag this number of phrases has encouraged many researchers like ourselves to create automatic applications for this issue. Our approach [28, 32] represents a completely automatic method for enlarging an already existing corpus, so it acquires the desired number of tagged phrases. The extra content of the corpus will be obtained from any knowledge source, such as the web, from which we extract untagged sentences to be analyzed. Considering the initial small corpus as the seed, our method makes it evolve until it reaches the target size in a totally automatic way.

Named entity recognition (NER). We are interested in identifying and labeling semantic information in a text, in such a way as to allow repeatable semantic patterns to emerge. So, we seek to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of time, quantities, monetary values or percentages. The more complex the ontology, the less accurate the classification, thus originating semantic and performance problems [2]. To alleviate this problem one can increase the size of the annotated corpus, but this is a very expensive task so we have designed a system [34] which provides a number of features which facilitate the visualization and tagging of annotated text. In the development of our NER system [7, 8], we have used the Spanish corpus distributed for the CoNLL02 shared task, and the tagger generator TnT. In order to improve performance, we have defined three transformations that give us additional versions of the training corpus, and we have trained TnT with them to obtain different taggers. Then we have applied a stacking scheme [9] to combine the results of the models.

Term extraction. In dealing with French we have considered Fastr[1] and Acabit[2]. The former is a term analyzer and a variant recognizer. Given that Fastr does not implement term

[1] http://www.limsi.fr/Individu/jacquemi/FASTR/
[2] http://www.sciences.univ-nantes.fr/info/perso/permanents/daille/index.html

extraction in Spanish, we have adapted the grammatical meta-rules and the set of tags used in French to Spanish, profiting from the syntactic similarity between these two languages. ACABIT is a term acquisition tool that takes as its input a formated text and returns an ordinated list of candidate terms. Given that ACABIT uses the FLEMM tagger, we have adapted the input interface in order to allow inputs from TREETAGGER. In the case of Spanish, we have also considered two different approaches. The first is a tool developed *ad-hoc* at the SYNTAX research group in the Univ. of Santiago de Compostela and it uses a TREETAGGER entry.

Spelling correction. We have also worked on the correction of the errors[3] present in the document collection, in order to make the error rate of the parser as low as possible. In this context, we avoid techniques traditionally implemented in commercial applications, based on a hand-made correction according to a list of candidate replacement options. We propose to apply fully automatic techniques, searching for repairs in a limited context which is dynamically determined for each error detected. We baptize this kind of error recovery technique as *regional* [17, 16, 20, 22, 21].

2.2 Syntactic level

Shallow parsing. This is a kind of "light parsing" where the analysis of a sentence seeks to identify the main constituents, also called chunks, but does not specify their internal structure, nor their role in the main sentence. We have developed a shallow parser, implemented by means of finite-state transducers with a 5-layer architecture [11, 12]. Once the parsing has finished, the syntactic function of each phrase detected is identified. Then, the dependencies between the heads of phrases are extracted [11, 12]. These dependencies are general enough to be present in any of the target languages of the project [10, 60, 59], taking into account that our work focuses on IR applications [62, 47, 61, 13]. Trying to make the technique as general as possible, we have also developed a technique for the compilation of parsing schemata [29, 66, 30], a declarative, high-level formalism for the description of parsing algorithms that can be used for any grammar in the Chomsky hierarchy [53, 54]. The goal is their application in the NLP domain [31] and, in particular, in the case of IR tools [55].

Robust parsing. This is the ability to find partial parses in an ungrammatical input, in such a way as to enable parsing to continue in the presence of parts of the input unexplained by the language grammar. Typically this involves producing a parse tree even in the presence of syntax errors, corrupted or unknown sentences. Our proposal is a regional least-cost strategy which applies a dynamic validation in order to avoid cascaded errors, gathering all relevant information in the context of the error location. The system guarantees the asymptotic equivalence with global repair strategies [15, 18].

2.3 Semantic level

Knowledge acquisition. We focus [52] on extracting and then connecting terms in order to detect pertinent relations and eliminate non-deterministic interpretations. To deal with, two principles are considered: the *distributional semantics model* establishing that words whose meaning is close often appear in similar syntactic contexts; and the assumption that terms shared by these contexts are usually nouns and adjectives. The parse takes the form of a graph

[3] typos, conversion errors, OCR errors, etc.

whose arcs represent relations of the type *governor/governed*, which permits the lexicon to be concentrated around *pivot terms* and even establishes similarity measures between these. Term extraction is organized around the recognition of generic lexical and/or syntactic patterns from these pivot terms. We profit from this topological information to apply automatic learning techniques in order to locate those dependencies that are more frequent and less ambiguous, focusing the meaning of the text on what we baptize as *strong dependencies*. At this point, we can infer a number of semantic tags that we use for text indexing. Also, once basic syntagmas and properties have emerged from text, we focus on more sophisticated patterns connecting them in order to derive more complex semantical relations as, for example, *hyperonymy*.

Parallel text alignment. Alignment is usually done by finding correspondence points. Early works used punctuation signs as alignment points, dividing texts into parallel sentences. Later, homograph tokens in both texts were used as correspondence points and enabled the identification of segments smaller than the sentence. Other methods have relied on using cognates, i.e. similar words having the same meaning as additional correspondence points. In particular, it is our aim to evaluate how cognates affect the quality of the results. For this purpose we used a simple measure of similarity which accepts as possible cognates all pairs whose similarity is above a certain threshold. Then, we studied [51] how the alignment is affected by changing this threshold. We have developed a methodology to assess the quality of resulting alignments, determining the most probable causes of misalignment and evaluating how these errors are affected by changes in our cognates' thresholds.

Ranking documents. We have also investigated [3, 27] how to adapt the TextRank method to make it work in a supervised way. TextRank is a graph based method that applies the ideas of the ranking algorithm used in Google (PageRank) to nlp tasks, such as text summarization, keyword extraction or word sense disambiguation. TextRank operates in an unsupervised way, without using any training corpus. Our main contribution is the definition of a method that allows us to apply TextRank to a graph that includes information generated from a training tagged corpus.

2.4 Indexing models

A syntactic-dependency based model. A model based on the use of the syntactic dependencies extracted by the shallow parsing, jointly with single terms, has been developed. We have considered the use of syntactic dependency pairs obtained from the topic and the use of syntactic dependency pairs obtained from top documents retrieved. Our experiments [10] indicate that the use of syntactic information is useful for refining the results obtained through single-terms, but improvement is limited by the noise introduced by syntactic dependencies when they are not accurately selected.

A pseudo-syntactic model. A pseudo-syntactic approach was also tested [11, 12, 62]: the distance between terms appears as a practical alternative that avoids problems arising from a grammar or a parser, and is also language-independent. In addition, the information about the occurrence of individual words can be integrated in a consistent way with the information about proximity to other terms, which in turn is often related with the existence of syntactic relations between such terms. Then, we have applied a locality-based model that considers the collection to be indexed not as a set of documents, but as a sequence of words where each occurrence of a query term has an influence on the surrounding terms. Such influences are additive, thus, the contributions of different occurrences of query terms are summed, yielding

a similarity measure. As a result, those areas of the texts with a higher density of query terms, or with important query terms, show peaks in the resulting graph, highlighting those positions of the text which are potentially relevant with respect to the query. It is worth noting that relevant portions are identified without the need to perform any kind of splitting in the documents, as is done in passage retrieval. Our experiments on this approach did not improve the results obtained by the syntactic-dependency based model.

A data-fusion based model. We have also merged the two previous approaches, using data fusion [61, 47, 13], a technique of combination of evidences that consists of combining the results retrieved by different representations of queries or documents, or by different retrieval techniques. In our approach, we have opted for using the following boolean criterion: First, we retrieve the documents contained in the intersection of the top documents retrieved by both approaches. Next, the documents retrieved in the top documents by only one of the approaches are retrieved. Finally, the rest of documents retrieved using single-terms. In our experiments, the improvements attained with this new approach were general, particularly in the case of precision at N documents, without penalizations for non-interpolated and R-precision.

Cross-lingual retrieval. We have developed [60, 59] an n-gram model that takes as input previously existing aligned word lists and obtains as output aligned n-gram lists. It can also handle word translation probabilities, as in the case of statistical word alignments. This solution avoids the need for word normalization during indexing or translation, and it can also deal with out-of-vocabulary words. Since it does not rely on language-specific processing, it can be applied to very different languages, even when linguistic information and resources are scarce or unavailable. Our proposal adds a high speed during the n-gram alignment process to these characteristics.

A sentence matching based model. We exploit the meaning of single-terms by integrating them into an edit distance construction. So, we extend a matching algorithm in a similarity parse tree measure taking into account the semantic proximity between words [14, 5, 6, 56]. This allows us to deal with similarity problems where we can take advantage of the use of semantic information in pattern-matching processes. We have chosen to compute a modified edit distance where the data tree can be simplified by removing some irrelevant subtrees with no associated cost. We also support the use of *variable length don't care* (VLDC) symbols in the pattern tree, which allow us to omit structural details and manage more general patterns. To formalize semantic similarity at word level, we use the WordNet taxonomy. On this basis we compute the semantic cost associated with the edit operations applied to the words in the sentences we are comparing. Semantic and syntactic distances are computed in parallel, propagating the similarity measure at word level through the nodes in accordance with the syntactic distances computed by tree matching.

3 Result Indicators

3.1 Formation Activities

During the project, J. Vilares (UDC) has finished the Ph.D. Thesis entitled *Application of Natural Language Processing to Spanish Information Retrieval*. Co-directed by M.A. Alonso (UDC) and J.L. Freire (UDC), 2005. The thesis was awarded the highest mark of "Sobresaliente Cum Laude por Unanimidad", the European Doctor Mention and the UDC Doctorate Prize.

The examination committee was formed by Gabriel Pereira Lopes (New University of Lisbon, Portugal), John I. Tait (University of Sunderland, UK), Éric de la Clergerie (INRIA, France), C. Martín Vide (Rovira i Virgili University, Spain) and J. Graña Gil (UDC, Spain).

The following Ph.D. Thesis are expected to be finished towards the end of the project:

1. F.M. Barcala (UVIGO). Co-directed by J. Graña (UDC) and M. Vilares (UVIGO), he works on improving the management of linguistic information for practical applications by developing specific IR algorithms and techniques for structured collections.
2. C. Gómez (UDC), recipient of a FPU fellowship from the MEC since May 2006, is also working on his Ph.D. thesis, with the aim of designing a general technique for generating robust parsers from declarative specifications in order to build practical natural language tools applicable to the fields of IR and QA. He has joined the project in May 30, 2006. The thesis is being directed by M.A. Alonso (UDC) and M. Vilares (UVIGO).
3. J. Otero (UVIGO), recipient of a FPI fellowship from the MEC since 2005, is working on his Ph.D. thesis under the direction of J. Graña (UDC) and M. Vilares (UVIGO), with the aim of designing a general technique for generating robust spelling correction techniques.
4. M.A. Molinero (UVIGO) has presented his DEA (UDC), and is recipient of a pre-doctoral fellowship from the Xunta de Galicia. He is working on the application of complex lexical and syntactic information to IR. This thesis is being directed by M. Vilares (UVIGO).

Additionally, M. Fernández (UVIGO), C. Gómez (UVIGO, currently FPU recipient in UDC) and S. Carrera (UVIGO) have been hired to work on the project and they will develop their Ph.D. Theses on subjects related to it. Finally, F. Enríquez (USE), F. Cruz (USE) and C. Gómez (UDC) will present their DEAs this year from the results of the project.

3.2 Publications

Articles in journals indexed by ISI JCR

[1] F.M. Barcala, M.A. Molinero, and E. Domínguez. Information retrieval and large text structured corpora. *Lecture Notes in Computer Science*, 3643:91–100, 2005.

[2] J.M. Cañete and F.J. Galán. Towards a theory on the role of ontologies in software engineering problem solving. *Lecture Notes in Computer Science*, 3442:205–219, 2005.

[3] F. Cruz, J.A. Troyano, and F. Enríquez. Supervised textrank. *Lecture Notes in Artificial Intelligence*, 4139:632–639, 2006.

[4] E. Méndez, J. Vilares, and D. Cabrero. COLE experiments at QACLEF 2004 spanish monolingual track. *Lecture Notes in Computer Science*, 3491:544–551, 2005.

[5] F.J. Ribadas, M. Vilares, and J. Vilares. Semantic similarity between sentences through approximate tree matching. *Lecture Notes in Computer Science*, 3523:638–646, 2005.

[6] F.J. Ribadas, J. Vilares, and M.A. Alonso. Integrating syntactic information by means of data fusion techniques. *Lecture Notes in Computer Science*, 3643:169–178, 2005.

[7] J.A. Troyano, V. Carrillo, F. Enríquez, and F.J. Galán. Named entity recognition through corpus transformation and system combination. *Lecture Notes in Computer Science*, 3230:255–266, 2004.

[8] J.A. Troyano, V.J. Díaz, F. Enríquez, and L. Romero. Improving the performance of a named entity extractor by applying a stacking scheme. *Lecture Notes in Computer Science*, 3315:295–304, 2004.

[9] J.A. Troyano, V.J. Díaz, F. Enríquez, and V. Carrillo. Applying stacking and corpus transformation to a chuncking task. *Lecture Notes in Computer Science*, 3643:150–158, 2005.

[10] J. Vilares, M.A. Alonso, and F.J. Ribadas. COLE experiments at CLEF 2003 Spanish monolingual track. *Lecture Notes in Computer Science*, 3237:345–357, 2004.

[11] J. Vilares and M.A. Alonso. Dealing with syntactic variation through a locality-based approach. *Lecture Notes in Computer Science*, 3246:255–266, 2004.

[12] J. Vilares, M.A. Alonso, and M. Vilares. Morphological and syntactic processing for text retrieval. *Lecture Notes in Computer Science*, 3180:371–380, 2004.

[13] J. Vilares, M.P. Oakes, and J.I. Tait. A first approach to CLIR using character n-grams alignment. *Lecture Notes in Computer Science*, 2006.

[14] M. Vilares, F.J. Ribadas, and J. Vilares. Phrase similarity through the edit distance. *Lecture Notes in Computer Science*, 3180:306–317, 2004.

[15] M. Vilares, V.M. Darriba, and J. Vilares. Parsing incomplete sentences revisited. *Lecture Notes in Computer Science*, 3945:102–111, 2004.

[16] M. Vilares, J. Otero, and J. Graña. On asymptotic finite-state error repair. *Lecture Notes in Computer Science*, 3246:271–272, 2004.

[17] M. Vilares, J. Otero, and J. Graña. Automatic Spelling Correction in Galician. *Lecture Notes in Computer Science*, 3230:51–57, 2004.

[18] M. Vilares, V.M. Darriba, J. Vilares, and F.J. Ribadas. A formal frame for robust parsing. *Theoretical Computer Science*, 328:171–186, 2004.

[19] M. Vilares, J. Otero, and J. Graña. Spelling correction on technical documents. *Lecture Notes in Computer Science*, 3643:131–139, 2005.

[20] M. Vilares, J. Otero, and J. Graña. Regional finite-state error repair. *Lecture Notes in Computer Science*, 3317:269–280, 2005.

[21] M. Vilares, J. Otero, and J. Graña. Robust spelling correction. *Lecture Notes in Computer Science*, 3845:319–328, 2006.

[22] M. Vilares, J. Otero, and J. Graña. Regional vs. global finite-state error repair. *Lecture Notes in Computer Science*, 3317:120–131, 2005.

[23] M. Vilares, J. Otero, and V.M. Darriba. Regional vs. global robust spelling correction. *Lecture Notes in Computer Science*, 3878:575–586, 2006.

Articles in other journals and books

[24] M.A. Alonso, E. de la Clergerie, V.J. Díaz, and M. Vilares. Relating tabular parsing algorithms for LIG and TAG. in Harry Bunt, John Carrol and Giorgio Satta (eds.), *New Developments in Parsing Technology* volume 23 in *Text, Speech and Language Technology Series*, chapter 8, pp. 157–184. Kluwer Academic Publishers, 2004.

[25] M.A. Alonso, J. Vilares, and F.J. Ribadas. Experiencias del grupo COLE en la aplicación de técnicas de procesamiento del lenguaje natural a la recuperación de información en español. *Inteligencia Artificial*, 22:123–134, 2004.

[26] F.M. Barcala, M.A. Molinero, and E. Domínguez. Construcción de sistemas de recuperación de información sobre córpora textuales estructurados de grandes dimensiones. *Procesamiento del Languaje Natural*, 34:41–48, 2005.

[27] F. Cruz, J.A. Troyano, F. Enríquez, and F.J. Ortega. Textrank como motor de aprendizaje en tareas de etiquetado. *Procesamiento del Lenguaje Natural*, 37:33–40, 2006.

[28] F. Cruz, J.A. Troyano, F. Enríquez, and F.J. Ortega. Ampliación automática de corpus mediante la colaboración de varios etiquetadores. *Procesamiento del Lenguaje Natural*, 37:11–18, 2006.

[29] C. Gómez-Rodríguez, J. Vilares, and M.A. Alonso. Generación automática de analizadores sintácticos a partir de esquemas de análisis. *Procesamiento del Lenguaje Natural*, 35:401–408, 2005.

[30] C. Gómez-Rodríguez, M.A. Alonso, and M. Vilares. Estudio comparativo del rendimiento de analizadores sintácticos para gramáticas de adjunción de árboles. *Procesamiento del Lenguaje Natural*, 37:179–186, 2006.

[31] C. Gómez-Rodríguez, J. Vilares, and M.A. Alonso. Automatic generation of natural language parsers from declarative specifications. *Frontiers in Artificial Intelligence and Applications*, 112:259–260, 2006.

[32] J. González, D. González, and J. A. Troyano. Adaptación del método de etiquetado no supervisado tbl. *Procesamiento del Lenguaje Natural*, 37:5–10, 2006.

[33] D. Guiziou, L. Félix, J. A. Gallegos, E. Sánchez, and Cr. Valderrey. *Fisioterapia: Glosario de convenciones textuales*. 2004.

[34] F. Ortega, V. Díaz, and L. Romero. Named: Una herramienta para la edición y manipulación de corpus. *Procesamiento del Lenguaje Natural*, 37:365–366, 2006.

[35] L. Rodríguez and C. García. Application of fusion techniques to speaker authentication over IP networks. IEE *Proceedings - Vision, Image and Signal Processing, I*, 2004.

[36] E. Sánchez. *Teoría de la traducción: convergencias y divergencias*. Servicio de Publicaciones de la Univ. de Vigo, 2004.

[37] E. Sánchez. Traducción de textos médicos entre el francés y el español: creación y explotación de corpus electrónicos. *Anales de Filología Francesa*, 12:395–412, 2004.

[38] E. Sánchez. Traducción de textos médicos: elaboración de un corpus electrónico francés-español. *Le français face aux défis actuels. Histoire, langue et culture*, 7:425-438, ISBN: 84-338-3238-7. 2004.

[39] E. Sánchez. Aproximación traductológica al análisis de corpus para el estudio de las convenciones textuales. In *Panorama actual de la investigación sobre traducción e interpretación*, pp. 113–127. Editorial Atrio, Granada, Spain, 2004.

[40] E. Sánchez. Corpus electrónicos y traducción: aplicación de las nuevas tecnologías a la traducción de textos médicos (francés-español). In *Insights into Scientific and Technical Translation*, pp. 267–273. Univ. Pompeu Fabre, Granada, Spain, 2004.

[41] E. Sánchez. Investigación traductológica en la traducción científica y técnica. *Revista de traductología*, 9:131–150, 2005.

[42] E. Sánchez. Traducción: de la teoría a las aplicaciones prácticas. In *Estudios de traducción, lingüística y filología dedicados a Valentín García Yebra*, pp. 203–218. 2005.

[43] E. Sánchez. Multidimensionalidad de la traducción científica y técnica entre el francés y el español. In *Manual de Traducción científica y técnica*. 2006.

[44] E. Sánchez. MYOCOR: creación y explotación de un corpus bilingüe (FR-ES) sobre enfermedades neuromusculares. *Revista de Traduçao Científica e Técnica*, 4:67–83, 2006.

[45] A. Sobrino, S. Fernández, and J. Graña. Access to a large dictionary of spanish synonyms: a tool for fuzzy information retrieval. *Studies in Fuzziness and Soft Computing*, 197:299–316, 2006.

[46] J. Vilares, C. Gómez-Rodríguez, and M.A. Alonso. Enfoques sintáctico y pseudo-sintáctico para la recuperación de información en español. In *Recuperación de información textual: aspectos lógicos y ecológicos — Text Information Retrieval: Soft-Computing and Ecological Aspects*, pp. 97–104. Servizo de Publicacións e Intercambio Científico, Univ. de Santiago, Santiago de Compostela, Spain, 2006.

[47] J. Vilares and M.A. Alonso. Tratamiento de la variación sintáctica mediante un modelo de recuperación basado en localidad. *Procesamiento del Lenguaje Natural*, 36, 2006.

Proceedings of international conferences

[49] F.M. Barcala, M.A. Molinero, and E. Domínguez. Information retrieval and large text structured corpora. In *Proc. of Tenth Int. Conf. on Computer Aided Systems Theory*, pp. 55-57, 2005.

[50] F.M. Barcala, M.A. Molinero, and E. Domínguez. XML rules for enclitic segmentation. In *Proc. of Eleventh Int. Conf. on Computer Aided Systems Theory*, 2007.

[51] G.P. Lopes, V.M. Darriba and T. Ildefonso. Measuring the impact of cognates in parallel text alignment. In *Proc. of 12th Portuguese Conf. on Artificial Intelligence*, pp. 334–343, Covilha, Portugal, 2005. Universidade de Beira Interior.

[52] M. Fernández, E. de la Clergerie and M. Vilares. From text to knowledge. In *Proc. of Eleventh Int. Conf. on Computer Aided Systems Theory*, Spain, 2007.

[53] C. Gómez-Rodríguez, M.A. Alonso, and M. Vilares. Generating XTAG parsers from algebraic specifications. In *Proceedings of the 8th Int. Workshop on Tree Adjoining Grammar and Related Formalisms*, pp. 103–108, East Stroudsburg, PA, 2006. ACL.

[54] C. Gómez-Rodríguez, M.A. Alonso, and M. Vilares. On theoretical and practical complexity of TAG parsers. In *Proc. of the 11th conference on Formal Grammar*, pp. 61–75. Center for the Study of Language and Information, Stanford, 2006.

[55] C. Gómez-Rodríguez, M.A. Alonso, and M. Vilares. Generation of indexes for compiling efficient parsers from formal specifications. In *Proc. of Eleventh Int. Conf. on Computer Aided Systems Theory*, 2007.

[56] F.J. Ribadas, J. Vilares, and M.A. Alonso. Integrating syntactic information by means of data fusion techniques. In *Proc. of Tenth Int. Conf. on Computer Aided Systems Theory*, pp. 96–99, 2005.

[57] F.J. Ribadas, E. Lloves, and V.M. Darriba. Multiple label text categorization on a hierarchical thesaurus. In *Proc. of Eleventh Int. Conf. on Computer Aided Systems Theory*, 2007.

[58] E. Sánchez. Le corpus électronique MYOCOR: Application des nouvelles technologies à la traduction des textes médicaux. In *Actes du XVIIe Congrès mondial Fédération Int. des Traducteurs*, pp. 155–157, 2005.

[59] J. Vilares, M.P. Oakes, and J.I. Tait. CoLesIR at CLEF 2006: Rapid prototyping of a n-gram-based CLIR system. In Carol Peters Alessandro Nardi and José Luis Vicedo, editors, *Results of the CLEF 2006 Cross-Language System Evaluation Campaign, Abstracts of the* CLEF *2006 Workshop*, Alicante, Spain, 2006.

[60] J. Vilares, M.P. Oakes, and J.I. Tait. CoLesIR at CLEF 2006: Rapid prototyping of a n-gram-based CLIR system (abstract). In Carol Peters Alessandro Nardi and José Luis Vicedo, editors, *Results of the CLEF 2006 Cross-Language System Evaluation Campaign, Abstracts of the* CLEF *2006 Workshop*, Alicante, Spain, 2006.

[61] J. Vilares, C. Gómez-Rodríguez, and M.A. Alonso. Syntactic and pseudo-syntactic approaches for text retrieval. In Vicente P. Guerrero-Bote, editor, *First Int. Conf. on Multidisciplinary Information Sciences and Technologies*, pp. 104–108, Badajoz, Spain, 2006. Open Institute of Knowledge.

[62] J. Vilares, C. Gómez-Rodríguez, and M.A. Alonso. Managing syntactic variation in text retrieval. In Peter R.King, editor, *Proc. of the 2005* ACM *Symposium on Document Engineering*, pp. 162–164, New York, USA, 2005. ACM Press.

[63] M. Vilares, J. Otero and J. Graña. Spelling Correction on Technical Documents In *Proc. of Tenth Int. Conf. on Computer Aided Systems Theory*, pp. 73–76, 2005.

[64] M. Vilares, J. Otero and J. Vilares. Robust spelling correction. In *Proc. of Tenth Int. Conf. on Implementation and Application of Automata*, pp. 151-159, 2005.

Proceedings of national conferences

[65] E. Sánchez. Creación y explotación de recursos documentales sobre enfermedades neuromusculares. In *Actas* XXII *Congreso de* ASEM, pages 85–94, 2005.

[66] C. Gómez-Rodríguez, J. Vilares, and M.A. Alonso. Compilación eficiente de esquemas de análisis sintáctico. In Francisco Javier López Fraguas, editor, *Actas de* PROLE'05, pages 151–159, Madrid, Spain, 2005. Thomson Paraninfo.

3.3 Collaboration with other research groups

1. **The** ATOLL **research group,** INRIA **(France)**: Leading group in human language technology, directed by E. de la Clergerie.

   (a) E. de la Clergerie visited UDC and UVIGO from March 13 to 20, 2006. He gave two talks about "From meta-grammars to factorized grammars" and "Error mining in parsing output of large corpora". The aim was to create a larger consortium in order to present a project proposal for the VII Framework Programme.

   (b) In May 2005, E. de la Clergerie visited UDC to participate in the examination committee of the Ph.D. of J. Vilares.

   (c) E. de la Clergerie visited UVIGO from November 14 to November 16, 2004 to give the talk "DYALOG: A frame for robust treatment in NLP. Generation of meta-grammars. Knowledge acquisition from texts written in natural language".

(d) M. Fernández (UVigo) has done a 5-month post-graduate stay (from February 27 to July 31, 2006) and a visit (from November 22 to December 19, 2006) to work on methods and techniques for automatic knowledge acquisition from texts.

(e) F.J. Ribadas (UVigo) has done several post-doctoral stays (from May 31 to June 30, 2004; from November 18 to December 23, 2004; from November 23, 2005 to March 4, 2006) to work on the integration of linguistic knowledge in IR tools.

(f) In September 2004 and August/September 2005 M. Vilares (UVigo) visited the group with the aim of studying future collaborations in the domains of automatic knowledge acquisition and metagrammar-based parsers, respectively.

2. **The** GLiNt **research group, New Univ. of Lisbon (Portugal)**: Leading group in human language technology, directed by G. Pereira Lopes.

   (a) In May 2005, G. Pereira visited Udc to participate in the examination committee of the Ph.D. of J. Vilares.

   (b) G. Pereira visited UVigo from November 14 to November 16, 2004. He gave a talk about "Text alignment and its application to the development of machine translation tools".

   (c) V.M. Darriba (UVigo) has done a 5-month post-graduate stay (June 23-November 23, 2005) and a visit (January 7-February 18, 2005), to work on parallel-texts alignment.

   (d) J. Otero (UVigo) has done two 3-month post-graduate stay (March 1-May 31, 2006; September 16-December 16, 2005) to work on the integration of syntactic information in spelling correction tools.

3. **The Information Retrieval Group of the School of Computing and Technology, Univ. of Sunderland (UK)**: Leading group in human language technology, directed by Prof. J.I. Tait.

   (a) J. Vilares (Udc) has done a 5-month post-doctoral stay (from September 14, 2005 to February 15, 2006) with this group, working on methods and techniques for cross-language IR. He also gave a talk about "Managing Syntactic Variation in IR: Spanish as a Case in Point" in October 5, 2005. In late 2006, he again visited Tait's group for a 3-month post-doctotal stay.

   (b) In May 2005, J.I. Tait visited Udc to participate in the examination committee of the Ph.D. of J. Vilares.

   (c) M. Oakes visited Udc and UVigo from March 13 to 20, 2006. He gave a talk about "Regular sound-changes for Cross-Language Information Retrieval". The aim was to create a larger consortium in order to present a project proposal for the VII Framework Programme.

4. **The** I3S **Laboratory,** Cnrs**/Univ. of Nice**: Leading laboratory in computer languages technology. In September, 2006, M. Vilares (UVigo) visited the I3S Laboratory invited by Dr. Jacques Farre. The goal of this visit was the study of future collaborations with I3S and Inria in the domain of tree-adjoining grammar parsers.

5. **The Dept. of Computer Languages and Systems, Univ. of Alicante**: F. Cruz (USE) has started a collaboration with Paloma Moreda in the area of Semantic Role Labelling (SRL). We have already planned a visit for her to our department to help us in the beginning of our investigations.
6. **The Univ. of Rochester, NY (USA)**: F. Enríquez de Salamanca (USE) has received an invitation from Mr. James F. Allen (URCS Faculty Member) to spend an unspecified period of time working with his research group, in order to share knowledge and find tasks in which we could collaborate in the near future.
7. **The Group of Spanish Syntax, Univ. of Santiago de Compostela**: Leading group in computational linguistics, directed by Prof. Guillermo Rojo:
   (a) From September 11 to 12, 2006, C. Gómez (UDC) participated in a meeting with G. Pereira Lopes (New Univ. of Lisbon, Portugal) and Gaël Dias (Univ. of Beira Interior, Portugal). He gave a talk on "Compilation of parsing schemata"
   (b) This group participates with our groups in UDC and UVIGO, and also other research groups in the universities of Galicia in the thematic grid *Galician grid for language processing and information retrieval*; led by Prof. M. Vilares (UVIGO), 2006-2009.
   (c) F.M. Barcala has collaborated in the MEC project "Multilingual creation and integration of linguistic resources by terminological and discursive control strategies in specialized communication domains" (HUM2004-05658-C02-02).
8. **The Natural Language and Computational Linguistics Group, Univ. of Sussex (UK)**: Leading group in parsing and generation, directed by Prof. John Carrol. A 4-month stay of C. Gómez (UDC) has been planned to work on the development of QA strategies on TAGs.
9. **The Dept. of Computer Science and Engineering, Univ. of North Texas (USA)**: J.A. Troyano (USE) has been in touch with Rada Mihalcea, who is one of the original authors of the TEXTRANK algorithm. We have sent our main paper about an extension of this algorithm and her response has been very encouraging, and we are sure that her advice and revisions will be of great help in our research.
10. **Laboratory of Agent Modelling-LabMAg, Dept. of Computer Science, Univ. of Lisbon (Portugal)**: J. Vilares (UDC) visited this laboratory from May 29 to June 5, 2005. He gave a talk about "Dealing with Syntactic Variation in Spanish IR".
11. **The Information Retrieval Group, Dept. of Information Studies, Univ. of Sheffield (UK)**: On October 18, 2005, J. Vilares (UDC) visited it at the invitation of Profs. Mark Sanderson and Paul Clough, who have a great experience in the field of cross-lingual IR and take part in the organization of CLEF campaigns. J. Vilares gave a talk about "Parsing-based vs. locality-based approaches for European Text Retrieval".
12. **The Information Retrieval Group, Dept. of Computing Science, Univ. of Glasgow (UK)**: On November 14, 2005, J. Vilares (UDC) visited this group, invited by Prof. Keith van Rijsbergen. It is a leading group in IR, being the creators of the TERRIER system and winners of the 2005 edition of TREC. During his visit, Vilares gave a talk on the issue "From multiwords to distances: dealing with syntax in IR".

13. **The Question Answering Group, Univ. of Edinburgh (UK)**: On February 10, 2006, J. Vilares (UDC) visited this group, invited by its coordinator Prof. Bonnie Webber. This group has created the QED (Edinburgh's QA system) and the WEE/QAAM, a shallow web QA system based on information fusion techniques.
14. **The Research Group in Computational Linguistics, School of Humanities, Languages and Social Studies; Univ. of Wolverhampton (UK)**: On December 12, 2006, J. Vilares (UDC) visited this group, invited by its director Prof. Ruslan Mitkov, a leading researcher in anaphora resolution, whose other current research interests include centering, automatic abstracting and term extraction.
15. **Dept. of Translation at the Univ. of Granada**, directed by Prof. P. Faber who has visited UVIGO from March 13 to 20, 2006. He gave a talk about "Terminology oriented to process in coastal and port engineering".

3.4 Collaboration with companies and institutions

1. MISTER DOC, a Portuguese company in the domain of computer science technology applied to the development of information systems, has established a pre-protocol of collaboration with the research group of UVIGO in order to design IR applications.
2. TELÉMACO, an Spanish company in the domain of IR systems, has collaborated in the projects "Generating, extracting and structuring legal information by means of artificial intelligence techniques" (Xunta de Galicia PGIDIT05SIN044E, 2005-2006, led by F.J. Ribadas from UVIGO and J. Vilares from UDC) and "Tools for the automatic analysis and classification of documents in the legal domain" (Diputación de Ourense 2005-INOU-09, 2006, led by V.M. Darriba from UVIGO). A pre-protocol of collaboration with the groups of USE, UDC and UVIGO in order to design QA applications has been established.
3. 3.14 FINANCIAL CONTENTS, a Spanish company in the domain of accessing economic information, has collaborated in the projects "Who, what, where, when, how many, how much, how and why? NLP tools, machine learning and Bayesian networks to build question-answering robots in financial markets" (Xunta de Galicia PGIDIT05SI059E, 2005-2008, led by V.M. Darriba from UVIGO) and "Application of Artificial Intelligence for extracting cognitive and qualitative information from financial markets" (Xunta de Galicia PGIDIT02SIN01E, 2002-2005, led by M. Vilares from UVIGO).
4. INDISYS, an Spanish company in the domain of intelligent dialogue systems, has established a pre-protocol of collaboration with the research groups of USE, UDC and UVIGO in order to design QA applications.
5. LA VOZ DE GALICIA, a Spanish newspaper, has established a pre-protocol of collaboration with the research groups of USE, UDC and UVIGO in order to design QA applications.
6. ASEM, Spanish National Association for Neuromuscular Illness, has collaborated in the project "Creation and application of documentary resources about neuromuscular illness" (Xunta de Galicia PGIDIT04SIN065E, 2004-2007, led by E. Sánchez from UVIGO).
7. The Ramón Piñeiro Center for Research on Humanities (Xunta de Galicia) has been involved in the development of the project "Tagger-Lemmatizer for Current Galician" since 1996. The latest contract covered the period 2004 to 2006 with UVIGO (led by M. Vilares) and UDC (led by J. Graña).