PROGRAMA JUAN DE LA CIERVA
MEMORIA DE LA ACTIVIDAD DE INVESTIGACIÓN A REALIZAR POR EL INVESTIGADOR CANDIDATO DENTRO DEL EQUIPO DE INVESTIGACIÓN
(DESCRIPTION OF THE RESEARCH ACTIVITY TO BE CARRIED OUT BY THE RESEARCHER AS PART OF THE RESEARCH TEAM)

(se deberá adjuntar una memoria para cada uno de los investigadores candidatos incluidos en el apartado B (Relación de investigadores candidatos))

Indicar las tareas y responsabilidades que desarrollará el investigador candidato dentro del equipo de investigación.
(Indicate the tasks and responsibilities that the researcher will conduct as part of the research team)

CUMPLIMENTAR PREFERIBLEMENTE EN INGLÉS – FILL IN BETTER IN ENGLISH

In the research presented in this proposal I aim at developing methods for the automatic extraction of semantic information from large data sources, such as existing corpora or the Internet. Currently, Natural Language Processing (NLP) techniques are quite mature with respect to grammatical aspects of language: the techniques to build morphosyntactic taggers are well understood, and part-of-speech taggers for Indo-European languages have reached 95-96% accuracy; shallow parsers are customarily used for research as well as commercial purposes (Manning and Schütze 1999). Deep syntactic analysis based on the automatic induction of grammars from syntactically annotated corpora is also a well-established technique (Collins 1999, Charniak 2000), although it has a limited performance, which, for instance, does not go beyond 85% for Spanish (Cowan and Collins 2005).

All these aspects cover formal aspects of language, while for most research and commercial purposes, access to content is a much higher priority (Ide and Véronis 1998; Nirenburg and Raskin 2004), so that semantic processing is called for. Semantic information is, however, much more difficult to encode and process through automatic means than morphosyntactic information. A large part of the difficulty involves establishing an adequate set of categories to represent semantic information, despite efforts in the development of lexical resources such as WordNet (Fellbaum 1998) or ontologies (Nirenburg and Raskin 2004) to support language processing. To date, all large-scale lexical resources including semantic information have been manually built, which necessarily causes inconsistency and incompleteness in the coding (Ide and Véronis 1998).

These difficulties have motivated the emergence of Lexical Acquisition, a field within Computational Linguistics that seeks to induce linguistic knowledge from corpora and other knowledge sources. The idea is to automate the process of building linguistic resources as far as possible, using the implicit information about words that can be extracted from language data. Research in the acquisition of selectional preferences, thematic role assignments and diathesis alternations (Korhonen et al., 2000; McCarthy 2001; Merlo and Stevenson 2001; Mayol et al. 2005), domain information (Magnini and Cavaglia 2000), lexicosemantic relations between words (Agirre and Martinez 2001), predominant senses (McCarthy et al. 2004), semantic classification (Schulte im Walde 2006, Boleda et al. 2004), etc., has obtained encouraging results.

In previous research (Boleda et al. 2004, Boleda 2007, Boleda et al. to appear), I have followed this line of research, focusing on the semantic classification of Catalan adjectives. The research has aimed at integrating computational techniques and linguistic research, so that results obtained through machine learning experiments have contributed to a better understanding of the linguistic issues involved in the semantic analysis of adjectives. I have also carried out several experiments involving human subjects (the largest of which involved 322 subjects), which have provided further insight into the difficulties faced in building robust semantic resources.

In the present research, I intend to generalise the methodologies I have developed to tackle the semantics of Catalan adjectives, and those developed in the related research cited above, to other parts of speech and other languages. Specifically, I aim at acquiring subcategorisation and selectional preference information for verbs, nouns, and adjectives in Catalan and Spanish. I will model the linguistic behaviour of the target lexical items on the basis of information extracted from corpora and lexical resources. The classification experiments will be carried out with machine learning techniques, with special emphasis on building ensemble classifiers that can richly combine different types of linguistic information.

Furthermore, the acquired information will be integrated within semantic processing systems, for tasks like Semantic Role Labeling and Question Answering.

My research will contribute to the KNOW project by addressing several of its challenges:

a) **multilingualism**, by carrying out parallel acquisition experiments for Catalan and Spanish and integrating the knowledge gathered for these languages with the project resources for Basque;
b) **large-scale knowledge acquisition**, by focusing on the fully automatic induction of information;
c) **open-domain semantic processing**, by building and enhancing general-purpose resources and integrating them within larger semantic processing systems.

My experience in linguistic research, and specifically in the use of computational techniques for linguistic problems, will be useful to the project and to the GPLN group in general. I will be an active member of the group, participating in seminars and research activities, contributing to other projects, collaborating with researchers in joint publications, and taking part in competitive project proposals. On the other hand, my career as a researcher will greatly benefit from the interaction with the GPLN group. In particular, I will profit from their expertise in machine learning approaches to language and in a broad range of NLP tasks.

References

Agirre, E., D. Martinez. 2001. Learning class-to-class selectional preferences. *Proceedings of the Workshop on Computational Natural Language Learning (CoNLL)*. In conjunction with *ACL'2001/EACL'2001*. Toulouse, France.

Boleda, G. 2007. Automatic acquisition of semantic classes for adjectives. PhD Thesis. Pompeu Fabra University, Spain.

Boleda, G., T. Badia, E. Batlle. 2004. Acquisition of Semantic Classes for Adjectives from Distributional Evidence. *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pp. 1119-1125, Geneva, Switzerland.

Boleda, G., S. Schulte im Walde, T. Badia. To appear. Analysis of Agreement about the Semantic Class of Adjectives. Accepted for publication in *Research on Language and Computation: special issue on ambiguity and semantic judgments*, edited by Massimo Poesio and Ron Artstein.

Charniak, E. 2000. Maximum-Entropy-Inspired Parser. *Proceedings of the North American Chapter of the Association for Computational Linguistics Conference (NAACL)*. Seattle, Washington.

Collins, M.. 1999. Head-Driven Statistical Models for Natural Language Parsing. PhD Dissertation, University of Pennsylvania, USA.

Cowan, B., M. Collins. 2005. Morphology and Re-ranking for the Statistical Parsing of Spanish. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Vancouver, Canada.

Fellbaum, C. (Ed.). 1998. *WordNet: an electronic lexical database*. London: MIT.

Ide, N., Véronis, J. 1998. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24(1), 1-40.

Korhonen, A., G. Gorrell, D. McCarthy. 2000. Statistical Filtering and Subcategorization Frame Acquisition. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. Hong Kong, Japan.

McCarthy, D. 2001. Lexical Acqusition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences. PhD Thesis. University of Sussex, UK.

McCarthy, D., Koeling, R., Weeds, J., Carroll, J. 2004. Finding predominant senses in untagged text. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Barcelona, Spain, pp. 280-287.

Magnini B., G. Cavaglia. 2000. Integrating Subject Field Codes into WordNet. In Gavrilidou, M., G. Crayannis, S. Markantonatou, S. Piperidis, G. Stainhaouer (Eds.), *Proceedings of LREC-2000 (Second International Conference on Language Resources and Evaluation.* Athens, Greece, pp. 1413-1418.

Manning, Christopher D., H. Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Mayol, L., G. Boleda, T. Badia. 2005. Automatic acquisition of syntactic verb classes with basic resources. *Language Resources and Evaluation*, 39(4), 295–312.

Merlo, P., S. Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3), 373–408.

Nirenburg, S. and Raskin, V. (2004). *Ontological Semantics*. Cambridge, MA: MIT Press.

Schulte imWalde, S. (2006). Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, 32(2), 159–194.