International Workshop
on Spanish Language Processing Technologies
Santa Fe, July 1997

The ARIES toolbox: a continuing R+D effort

Authors: José C. González
José M. Goñi
Amalio F. Nieto
Antonio Moreno†

Institution: Universidad Politécnica de Madrid
†Universidad Autónoma de Madrid

Contact: José C. González
E.T.S.I. Telecomunicación
Ciudad Universitaria
E-28040 Madrid (Spain)

Telephone: +34 1 336 7299
FAX: +34 1 336 7333
E-mail: jcg@gsi.dit.upm.es

Proposal for Panel 3: Research and Development Infrastructure

The ARIES toolbox: a continuing R+D effort*

José C. González, José M. Goñi, Amalio F. Nieto and Antonio Moreno†
E.T.S.I. Telecomunicación
Universidad Politécnica de Madrid, Spain
†Facultad de Filosofía y Letras
Universidad Autónoma de Madrid, Spain

Abstract

The effort under the ARIES toolbox spans through the last six years. The core of the toolbox is its lexical platform, including a large Spanish lexicon, lexical maintenance and access tools and morphological analyser/generator. Upon this platform a set of tools have been implemented, including tokenizers, spell checker, unification-based parser and grammar, stochastic and neural morphosyntactic taggers, etc. On the side of applications, the current work is oriented towards offering networking linguistic services for the publishing industry.

1 Introduction

The work presented here has been carried out during the last six years to develop (almost from the scratch) a lexical platform both, portable and efficient (in terms of linguistic coverage of the morpho-syntactic phenomena, storage and access time). Upon this platform, a set of tools has been developed[1]. The main significance of these tools comes from the fact that they have been developed with a strong practical orientation and, simultaneously, starting from formalized linguistic knowledge.

2 Architecture of the lexical platform

The architecture of our lexical platform distinguishes two representation levels:

Source level: this is the human-readable level. It is designed to permit the encoding of linguistic generalizations through default inheritance, rules to compute morpheme allomorphs, and the grouping of related entries in lemmas.

*Until now, this work has been supported in part by the Spanish *Plan Nacional de I+D*, through the projects TIC91-0217C02-01, *An Architecture for Natural Language Interfaces with User Modelling*, and TEL96-1367, *ATILA : Aplicaciones Telemáticas de Ingeniería Lingüística*, and by the Commission of the European Communities under contract MLAP-93/20, *CRATER: Corpus Resources and Terminology Extraction.*

[1]Licenses for some of the components of the ARIES tools are offered. Follow this link for more information: http://www.mat.upm.es/~aries.

Object level: this is the computer-readable level. It is designed for the efficient retrieval of the entries by the applications. Different object dictionaries suited for different applications can be automatically compiled from the source level.

A formalism for source level lexical representation has been designed [Goñi et al., 1995, Goñi et al., 1997]. Its main features are expressiveness, versatility, economy of expression and non redundancy. The design of this lexical representation language was influenced by the strong reliance of the Spanish language on inflectional morphology (e.g. 53 simple word forms for verbs, up to 4 forms for nouns and adjectives). We adopted a computational model for the treatment of Spanish inflectional morphology that is described in [Moreno, 1992] and in [Moreno and Goñi, 1995]. The main features of this model are:

- Morphological processing is constrained to morpheme concatenation, so its allomorphic variants have to be stored or computed.
- The model follows a *Graphical Word criterion*, that only considers relevant its written form. This criterion requires that additional allomorphs be necessary in some cases, because of diacritical marks, or different surface realization of the same phoneme in different contexts.
- Models for verbs and nominals are described in order to capture some interesting and well founded linguistic generalizations in the inflectional behaviour of the Spanish language. Obviously, some lexicalized forms are also included for the case of highly irregular word forms not belonging to any of the proposed models.

Our lexical representation language includes devices to implement this morphological model:

- Inflectional models are encoded into classes, so information about the behaviour of the word forms can be inherited by lemmas. The inheritance mechanism implemented is multiple default inheritance, with a priority scheme.
- The different allomorphs needed by each lemma are not directly stored, but computed by means of regular expression-based rules that can be called by the classes or directly by the lemma entries.
- Sections of the lexical source base are designed for different types of entries, like lexicalized words (strong irregularities, invariable entries, etc.), morphemes and lemmas.
- A data dictionary section is included for the declaration of the possible feature names and values that can be attached to entries. This allows some limited type checking and efficient encoding of the compiled dictionary.

The ARIES source lexical base has now a considerable size (around 38,000 lemma entries), accounting for more than 400,000 inflected forms. It currently includes 76 classes (both inflectional and auxiliary ones), 53 allomorphy rules (with a total of 176 productions), 622 morphemes, 697 lexicalized invariable entries, 21,000 nouns, 10,000 adjectives and 7,300 verbs.

3 The ARIES tools

1. Dictionary compiler
   Object dictionary generation is achieved by specialized tools that are guided by filtering rules. These are a set of rules of a tree-manipulating language that allows the filtering, addition or modification of the branches of the feature bundle attached to each entry, as well as the computing of the different allomorphs for each lemma and their assignment of the relevant feature bundle. The tools (all of them implemented in C/C++) deal with the interpretation of filtering rules, inheritance attachment, interpretation of regular expression rules to compute the allomorphs, and building the *letter-tree* index (*trie*) for efficient retrieval.

2. Dictionary interface
   A flexible software access interface has been implemented to allow the applications a fast retrieval of lexical entries. A *trie* index is built while the object dictionary is compiled. This indexing mechanism has allowed us the integration in this interface of the word segmentation needed for morphological analysis. Thus, word segmentation is no longer a blind process, since it is guided by the entries included in the lexicon.

3. Parsing tools
   Applications can be built in a modular way, picking up the modules that are best suited for them.

   (a) Unification modules
   Two modules were built for the processing of feature structures contained in PATR-II rules. The first one is a full unification module that allows real value sharing among features. The second one is a pseudo-unification module that treats feature bundles as trees, without value sharing. The implementation is based on the Tomita's pseudo-unification algorithm with some enhancements.

   (b) The modular chart parsing system
   Our modular chart parsing system, called NUCLEO, is a module that serve as a basis for the development of chart parsers. It is based on the CMCHART architecture proposed by Thompson, and includes facilities for implementing parsers with different searching strategies, rule invocation strategies, and top-down, bottom-up, or mixed strategies.

   (c) Tools for morphological analysis/generation
   The parser is built upon the NUCLEO system, and is integrated with the full unification module. The grammars have to be written in a PATR-II like format to feed the parser. A parser with the pseudo-unification module is also available if so desired. Although it is not implemented yet, a robust parser can be built in this parsing system, since charts can be used to recover partial results in incorrect analysis.

   A prototype of morphological analyser/generator (GRAMPAL) was also developed in PROLOG. It is, basically, a DCG grammar for word formation. The advantages of this prototype are declarativity and bidirectionality. The

lexicon needed for its operation is automatically derived from our source lexical base. It is fully described in [Moreno and Goñi, 1995].

4. The tokenizer
   The main goal of the tokenizer is to split a text into tokens, which may be simple words, adverbial compounds, numerals, dates, acronyms, proper nouns, text between quotes or parenthesis, beginning and end of sentences, punctuation marks, paragraph delimiters or strange words (those which contain both numeric and alphabetic characters, or non Spanish characters).

5. The stochastic tagger
   In the context of the CRATER project[2], the Xerox Tagger has been adapted to Spanish. The Xerox Tagger uses a statistical method for text tagging.

   One of the main advantages of the Xerox Tagger is that it does not need any tagged corpus; it is only necessary a lexicon with the more common words and their corresponding tags, and a suffix lexicon which stores the set of tags which may correspond to a particular suffix[3]. With these elements it is possible to train the Hidden Markov Model (HMM) in which the tagger is based. The modifications made to the Xerox Tagger in order to adapt it to Spanish have to do mainly with suffix handling and processing of clitics pronouns.

6. The verbal subcategorization tool
   Verbal subcategorization frames are the different syntactic structures that a particular verb can admit: types of complements (and required prepositions), non-finite forms, reflexive forms, auxiliaries, etc. A tool [Monedero et al., 1995] has been developed to automatically obtain the frames corresponding to Spanish verbs from tagged texts. 25 different subcategorization frames are considered, including simple, compound and auxiliary frames.

7. Demonstrators
   An e-mail interface has been built to show the capabilities of the tools developed until now. The address to which petitions can be sent is aries@mat.upm.es. Using a Web browser, the ARIES page can be reached at the URL http://www.mat.upm.es/-~aries, where forms for the e-mail interface are also available.

4 Work under development

At this moment, in the framework of the project ATILA, a set of linguistic distributed services are being developed for the publishing industry in general, and for the daily press in particular. In this work project, the Spanish newspaper EL PAIS, participates as user

[2]Our group contributes to this project as subcontractor of *Laboratorio de Lingüística Informática, Universidad Autónoma de Madrid, Spain*, with Fernando Sánchez being the project leader of the Spanish team.

[3]This suffix lexicon has been generated manually by Fernando Sánchez; the original tagger has the possibility of generating suffixes automatically from a corpus, but this method produces poor results in the case of Spanish.

and reviewer. The applications involved in the project are oriented towards the linguistic and style checking of texts and towards the field of information retrieval (indexing and retrieval of articles and photographs). During this year a JAVA interface is being built for the access to these distributed services, besides two demonstrators in the fields mentioned above.

Other topics we are currently working in are:

- Several versions of taggers, using neural and combined stochastic and knowledge-base approaches.
- Enhanced version of the ARIES lexicon.
- Phrase structure analyzer.
- Improved version of other tools: tokenizer, morpho-syntactic analyzer, etc.

Some publications from our group

[Goñi et al., 1995] Goñi, J.M. and González, J.C. A framework for lexical representation. *Proceedings of AI'95: Fifteenth International Conference. Language Engineering '95, pp. 243–252. Montpellier, June 27–30, 1995.*

[Goñi et al., 1997] Goñi, J.M.; González, J.C. and Moreno, A. (1997). ARIES: A Lexical Platform for Engineering Spanish Processing Tools. *To appear in the journal Natural Language Engineering.*

[Monedero et al., 1995] Monedero, J.; González, J.C.; Goñi, J.M.; Iglesias, C.A. and Nieto, A. (1995). Obtención automática de marcos de Subcategorización verbal a partir de texto etiquetado: el sistema SOAMAS. *XI Conferencia de la Sociedad Española para el Procesamiento de Lenguaje Natural, SEPLN-95, Bilbao, Septiembre, 1995.*

[Moreno, 1992] Moreno, A. Un modelo computacional basado en la unificación para el análisis y la generación de la morfología del español. *Tesis Doctoral. Universidad Autónoma de Madrid, 1992.*

[Moreno and Goñi, 1995] Moreno, A. and Goñi, J.M. GRAMPAL: A morphological model and processor for Spanish implemented in Prolog. *1995 Joint Conference on Declarative Programming (GULP-PRODE'95), Marina di Vietri (Salerno, Italy), September, 1995.*

[Sánchez and Nieto, 1995] Sánchez, F. and Nieto, A.F. Development of a Spanish Version of the Xerox Tagger. *Universidad Autónoma de Madrid, May, 1995.*

José C. González

Associate Professor at the School of Telecommunication Engineering, Technical University of Madrid, and Head of the Technical Committee of the Research Center for Multimedia Technologies and Applications (CITAM).

Active researcher in the field of Natural Language Engineering since 1990, leading the following projects funded by national and european research agencies:

- *An Architecture for Natural Language Interfaces with User Modelling*, project TIC91-0217C02-01, funded by the Spanish *Plan Nacional de I+D*.
- *CRATER: Corpus Resources and Terminology Extraction.*, project MLAP-93/20, funded by the Commission of the European Communities (subcontractor of Universidad Autónoma de Madrid).
- *ATILA : Aplicaciones Telemáticas de Ingeniería Lingüística*, project TEL96-1367, funded by the Spanish *Plan Nacional de I+D*.