RESLA 19 (2006), 179-199

MPRO-SPANISH: DEVELOPMENT AND EXPERIMENTS WITH A LINGUISTIC PARSER FOR SPANISH TEXTS

YAMILE RAMÍREZ SAFAR, JOHANN HALLER
Universidad de Saarland
MARIONA SABATÉ CARROVÉ
Universidad de Lleida

ABSTRACT. *This paper describes the main features and present results of MPRO-Spanish, a parser for morphological and syntactic analysis of unrestricted Spanish text developed at the IAI[1]. This parser makes direct use of X-phrase structure rules to handle a variety of patterns from derivational morphology and syntactic structure. Both analyses, morphological and syntactic, are realised by two subsequent modules. One module analyses and disambiguates the source words at morphological level while the other consists of a series of programs and a deterministic, procedural and explicit grammar. The article explains the main features of MPRO and resumes some of the experiments on some of its applications, some of which still being implemented like the monolingual and bilingual term extraction while others need further work like indexing. The results and applications obtained so far with simple and relatively complex sentences give us grounds to believe in its reliability.*

KEYWORDS. *Automatic analysis of Spanish texts, automatic morphosyntactic analysis, procedural explicit grammar, term extraction, indexing.*

RESUMEN. *En el siguiente artículo se describen las características principales de MPRO-Spanish, un analizador morfosintáctico para el español desarrollado en el IAI[2]. MPRO utiliza reglas basadas en la teoría de la X-barra, lo cual le permite analizar diferentes patrones de la morfología derivacional y de la estructura sintáctica. Tanto el análisis morfológico como el sintáctico son realizados por dos módulos; el primero se encarga de analizar morfológicamente y desambiguar las palabras de entrada; el segundo consiste en una serie de programas y una gramática determinística, procedural y explícita. El artículo explica los resultados de experimentos realizados con algunas aplicaciones de MPRO-Spanish. Varias de estas aplicaciones tienen continuidad actualmente –por ejemplo la extracción terminológica–, mientras que otras necesitan de mayor investigación, como la indización. Los resultados obtenidos hasta el presente mediante análisis de oraciones relativamente complejas en español nos hacen creer en su capacidad de suministrar análisis de calidad.*

PALABRAS CLAVE. *Análisis automático de textos españoles, análisis morfosintáctico, gramática procedural explícita, extracción terminológica, indización.*

1. Introduction and objectives

MPRO-Spanish is a tool for automatic linguistic analysis of general and specialised Spanish texts developed in the framework of a multilingual project carried out at the Institute for investigation on applied information (IAI) of the University of Saarland. The system consists of a series of subprograms, dictionaries and lexicons as well as an explicit and procedural grammar, which interact with one another from two modules and aim at carrying out a morphological and syntactic analysis. MPRO-Spanish is only one of the manifold puzzle-pieces of the MPRO multilingual project, which includes other languages such as German, English, French, Portuguese, Italian, Modern Greek, Esperanto, Dutch, Swedish, Russian and Bulgarian.

The goal of this article is to present MPRO-Spanish and the results of some preliminary experiments carried out with this analyzer for Spanish in areas such as term extraction and indexing. The article is organized into six sections. The second section deals with the state of the art, that is to say, with the approaches generally adopted to develop morphosyntactic analyzers and with some of the existing tools for Spanish. Section three provides a description of the analysis process and architecture of MPRO-Spanish, its modules for morphological and syntactic analyses and its deterministic, procedural and explicit grammar. In section four, we sum up the work done so far on the lexica to improve specifically the morphological analysis and consequently the syntactical one. Section five explores the results of some experiments with MPRO-Spanish done in term extraction and indexing. Finally, section six presents a summary of the article and the future research lines.

2. State of the art

To our knowledge, there exist many tools for the processing of Spanish which are based on a *morphosyntactic analysis*. The linguistic analysis such programs carry out though is either partial or its results can hardly be used for purposes other than those for what they were originally conceived, that is, they have been developed with very specific objectives in mind –for example, corpora tagging[3]. As a result, many of their capabilities are limited from their conception. This is not the case with MPRO-Spanish as it has been conceived as a program whose range of applications is open and wide. Below we will briefly sketch some of the features of other morphosyntactic parsers for Spanish and some of their differences compared to MPRO- (MPRO from now on).

Although nowadays it is becoming increasingly necessary to incorporate syntactic and semantic analyses in order to further analysis possibilities, as suggested by Rojo, "a text morphosyntactically tagged and lemmatized is just one more step on the road" (Rojo 2001: 1: our translation), the successful realisation of many applications is still based on correct morphosyntactic analysis. There is today a variety of methods aimed at improving the quality of such type of analyses. Among them, two stand out: the *rule-based symbolic* or *linguistic* method and the *statistics-based* or *stochastic* method. Some

symbolic methods are *finite-state* and *unification algorithms*; some stochastic methods include the *Hidden Markov Model* (*HMM*), *vectorial spaces* and *clustering*. There is a third model, a *hybrid* model, which combines both methods (Chanod et al. 2001: 2-7). The three of them will be roughly sketched out below.

2.1. *Stochastic approach*

There is a series of systems following the stochastic approach that carry out morphological analyses for diverse purposes. One of the areas of application of morphological analysis is corpus generation/creation. For Spanish, several systems have been developed for automatic extraction of information and corpora annotation which generally infer their grammars from specific corpuses that need to be analysed syntactically. One of these systems is the *corpus tagger*, described in Subirats and Ortega (2002). Unlike MPRO, the *corpus tagger* fails to carry out a complete morphological analysis and requires a large lexicon which supplies full and inflected forms and Spanish derivatives for recognition and analysis of forms. Likewise, MPRO does not carry out disambiguation and determination of relation between predicates and arguments in two phases, but in one, since it is during disambiguation that the relation between predicates and arguments is established. Finally, MPRO offers the results of the sentence analysed in its original form and its morphosyntactic analysis –plain or in form of a tree– which clearly shows the syntactic components of the sentence.

Another parser for Spanish is the one described by Aone and Hausman (1996), called the *Unsupervised Learning Part of Speech Tagger* proposed by Brill (1995) for the purposes of reducing word ambiguity. Compared to MPRO, the morphological analysis that this parser produces is rather elementary as it only uses a series of rules to eliminate or modify the word endings and find the word roots. Moreover, manual disambiguation of texts is a fundamental requirement, particularly at the outset, during the rule-learning process.

Another example of stochastic automatic grammatical taggers for Spanish which follow the stochastic approach is the one developed by the Group of Computational Linguistics of the Centre for Applied Linguistics of Santiago de Cuba. This tagger was designed according to the *Hidden Markov Model* (HMM) and is currently being used for the analysis of any text corpus (Ruiz 2000). The hierarchical semantic information – which in MPRO is still at its initial stages – and the treatment of unknown words are some advantages of this program in front of MPRO. Unfortunately, this parser was not available for a test during the time of our research. (aquí quite lo de la velocidad, pues no podemos mencionar la rapidez sin decir qué tan buenos son los resultados)

Another system for morphosyntactic disambiguation consists of the parsers and integrated grammars developed by the Group for Language Processing from the *Universidad Politécnica de Cataluña* (*UPC*). The program components are the morphological analyzer **MACO+** (*Morphological analyzer Corpus-Oriented*), the syntactic annotator **Relax**, the grammatical tagger **TreeTagger** and the parser **Tacat**.

MACO+ establishes the morphological tags for each word in a sentence. Then the tagged text is run against the morphosyntactic disambiguator **Relax**, which takes care of the word disambiguation according to a set of predetermined restrictions. Finally, the syntactic analyzer **Tacat** carries out the analysis of the text without any restriction. The user interface allows us to determine those program components that will interact during the generation of the analysis. Like MPRO, the program enables to view the results both as tree graphics or as plain chains. The overall results obtained by these tools tend to be rather reliable. However, like MPRO, it is not always possible to reconstruct the structure of the whole sentence but rather that of its constituents.

2.2. *Symbolic approach*

Although in the 90's complete analysis of sentences started to be abandoned as a linguistic method giving way to stochastic methods, there are today many current systems that do not use any statistical method to carry out morphosyntactic analyses (Chanod et al. 2001: 4). Examples of this are the parsers CREA and CATMORF. The latter, in spite of being an analyzer for Catalan texts, includes a prototypical Spanish parser.

The lexical analysis of the parser developed by the Spanish Royal Academy (*RAE*) may be similar to that of MPRO in that the inflection forms are not stored as lemmas in a dictionary, but are generated through rules based on the minimal information from the lexicon. The difference by MPRO is that our parser undertakes a complete morphosyntactic analysis by using the lexical information in the dictionaries and the rules in the file of inflected forms which also describe flat structures with attribute and value pairs. Unfortunately, it is not possible to have direct access to the results of the analysis from the RAE parser, as it has been created specifically for the annotated texts in CREA –Current Corpus of the Spanish Royal Academy– and CORDE –Diachronic Corpus of the Spanish Royal Academy–. The text corpus is the database itself. CREA only offers the user *tokens* for the study of words, their meanings and contexts, but it does not allow any morphological or syntactic studies, as the corpus has not been annotated for lemmas. For example, it does not allow to obtain all the forms of a noun, verb or adjective. If we enter the Spanish verb form *ser* (to be) searching for information on its different forms, moods or tenses –*es* (is-are)*, fue* (was-were)*, era* (was-were), etc.–, we will only get corpus examples where the form *ser* appears exactly as it is. Equally, the search after syntactic constructions such as NP + V + ATTR is not possible. This shortcoming may be caused by the fact that the system fails to give a complete linguistic analysis.

Badia et al. (1997a, 1997b) describe the Catalan tagger CATMORF, designed on a *two-level morphology*, whereby the morphological analysis results from the application of two-level rules and grammars of word unification. The program analyses any SGML-coded text and includes a 70,000-word lexicon, obtained semi-automatically from an electronic Catalan dictionary. The lexicon contains information on the lemmas, inflection paradigm of verbs, nouns and adjectives (Badia et al. 1997a: 25-26). Unlike CATMORPH though, MPRO carries out the morphological parsing in a unique stage.

Furthermore, the dictionary of morphemes together with the file of inflected forms caters for derivational, inflection and compounding processes for Spanish.

3. MPRO'S ANALYSIS AND ARCHITECTURE

MPRO's components were initially written and designed under Prolog. Nowadays they are available in C program and still being developed on Unix workstations[4]. The entries in the dictionaries and lexica resemble Prolog-terms. Since there are different calls to run the different components of the program, it is possible to make only a morphological analysis or the complete one, that is to say, the morphosyntactic analysis.

The following two subsections provide a description of the analysis process and architecture in MPRO.

3.1. *LESEN: Segmentation and Morphological Analysis*

The morphological analysis of MPRO consists of separating and establishing the difference between grammatical information and the form or word root. The entry text of LESEN must be written in ASCII with no specific marks or it must contain SGML or HTML signs. During the morphological analysis, MPRO supplies a packet of information or *feature bundle* to a full succession of characters known as a *lexical unit* (*lu*). As a result of the morphological analysis, we obtain the information contained in the dictionary of morphemes on the basis of the word root or class –noun, verb, adjective, etc.–, as well as the information on inflection –number, gender, tense and person– and the internal structure of the analyzed word, that is to say, its constituent parts –if the word contains prefixes or is a compound– (Maas 1998: 168).

To begin with, MPRO identifies a string of characters or words separated by spaces or punctuation signs with a view to establishing the limits of the sentence. In order to do that, the program uses the information contained in a file (called **limitrules**), which establishes the type of signs that can be considered as marks to delimit a string of characters or sentence. This file also contains the rules needed to identify dates, titles, figures, e-mail addresses, etc. After identifying the limits of a string of characters, the program looks for all forms in that string that could be in the **dictionary of frequent words**. This dictionary includes only forms and words belonging to closed categories in Spanish –i.e. prepositions, articles, pronouns, conjunctions and adjectival and adverbial phrases– and it has approximately 1,280 entries. For recognition of proper nouns, surnames, toponyms, abbreviations and acronyms in Spanish, we have created a **dictionary of proper nouns**, similar to the **dictionary of frequencies** in that it includes only functional words which do not require any morphological analysis whatsoever and do not constitute morphological steams. This dictionary contains approximately 2,700 entries.

When a word cannot be found in any of these dictionaries, the program determines that it needs to undertake morphological analysis and this is when the module for morphological analysis LESEN comes into play. To minimize the number of morphological

decompositions and, therefore, the duration of the morphological analysis, LESEN uses the information contained in a file (called **wrong**), that includes approximately 80 Spanish alomorphemes whose morphological parsing is not possible. This helps to avoid, for example, the segmentation of words such as *la|tin* in *latinoamericano*.

Once the morphemes that cannot be segmented are identified, LESEN initiates the morphological analysis with a **dictionary of morphemes** and a **file of inflected forms**. The Spanish morphemes and alomorphemes can be found in the **dictionary of morphemes**, which contains approximately 40,000 entries. Every entry is coded according to its grammatical category with information such as grammatical category, gender, number, possible derived forms, ending, inflection, etc. Likewise, it is indicated if the form is reflexive (in the case of a verb or a pronoun) or if it belongs to one of the predefined semantic category (i.e. *process*, *agent*, *science*, *state*, *disease*). The morphological information and the value of the lemmas are represented through chains of attribute and value pairs. The following example shows the entry for the regular verb *diferenciar (*to differentiate*)* – with its syntactic category (*c=v*), the inflection group to which it belongs (*t=ar*), a possible nominal derivation (*n={ación=process}*) and its reflexive use (*rfl=yes*):

{string=diferenci, **c=v**, lu=diferenciar, **t=ar**, **n={ación=process}**, **rfl=yes**}

The following example corresponds to the noun entry *acompañante* (traducción en inglés) from the **dictionary of morphemes**, which is coded according with its syntactic category (*c=n*), its plural form (t=s) – which is build only with an "s" –, its genre (*g=m*) and semantic class (*s=agent*):

{string=acompañante,**c=n**,lu=acompañante,**t=s**,**g=m**,**s=agent**}

The **file of inflected forms** contains the necessary rules for word formation. In this file, all possible values for all the features present in every entry in the dictionary of morphemes is defined. In other words, this file contains the inflected or segmented forms for the Spanish morphology with their corresponding linguistic information, such as number, gender, conjugation, mood, tense, among others. With the rules in this file it is possible then to analyzed and generate automatically the possible forms of the roots contained in the dictionary of morphemes. To illustrate how this is possible, we take from the dictionary of morphemes the same entry as above for the verb *diferenciar* and some rules from the file of inflected forms which stand for the recognition and generation of some of the possible inflections of *diferenciar*:

{string=diferenci, c=v, lu=diferenciar, **t=ar**, n={ación=process}, rfl=yes}

Diferenciar is coded in the dictionary of morphemes with the ending t=ar, which means that the verb belongs to the regular Spanish verb class ending in *–ar*. Now, the rules to analyze and generate forms such as *diferencia* (he/she differentiates) and *diferenciaba* (I/he/she differentiated) – just to mention a few – are contain in the file of

inflected forms. The values for "t" in the dictionary of morphemes correspond to those of "l" in the file of inflected forms, in our example *t=ar* and *l=ar*:

a) {**string=ab**,**mo=áb**,ac=b,**c=infix**,**l=ar**;**ar-voc**;**fragu**,m={**t=imp**}}: This rule stands for the analysis and generation of the morpheme *ab* (*string=ab*), with or without accent (*mo=áb*) in *diferenci**ab**a* or *diferenci**áb**amos*. According to the information in the feature bundle, this morpheme is an infix (*c=infix*) of the verb classes *ar*, *ar-voc* and *fragu* and indicates imperfect (*t=imp*).

b) {**string=a**,**c=flex**,**l=ar**;**ar-voc**;**fragu**;**piens**,m={**vtyp=fiv**,**tns=pres**,**mode=ind**,**per=3**,**nb=sg**}}: With this rule it is possible to analyze and generate the morpheme "a" (*string=a*) as a verbal inflection (*c=flex*) of the verb classes *ar*, *ar-voc*, *fragu* and *piens*. The information this inflection form adds to the word is third person (*per=3*), present time (*tns=pres*), indicative mode (*mode=ind*) and singular (*nb=sg*). The analysis of forms like *diferenci**a*** and also *diferenciab**a*** are completed with this rule.

Every word or form morphologically analyzed receives not only packets of linguistic features, but also a set of specific "administrative information" such as number of the word in the text (*wnra*), number of the word in the sentence (*wnrr*), number of the sentence in the text (*snr*) and so on. Whenever the program creates a complete structure of all words in the text or sentence, the morphological analysis is considered to be finished and results are shown. After that, no more morphological rules are applied.

The example below from the file *scr.les* –where words are not yet disambiguated– shows the morphological analysis obtained with MPRO for the following nominal phrase "*la reforma del paro*" (The reform of the system of unemployment benefits):

{wnra=15,wnrr=11,snr=2,**ori=la**,pctr=no,last=no,pctl=no,offset=85,lw=no,gra=small,**c=d**,subj=no,obj=acc,gov=no,fu=np,intrel=no,negated=no,poss=no,prp=no,pred=no,nb=sg,subl={reg=0,dom=0,loc=no,freq=0},ds=la,ls=la,w=1,s=nil,ew=0,lu=la,saw=&b}

{wnra=15,wnrr=11,snr=2,**ori=la**,pctr=no,last=no,pctl=no,offset=85,lw=no,gra=small,**c=w**,**sc=art**,nb=sg,g=f,fu=a,subl={reg=0,dom=0,loc=no,freq=0},ds=art_b,ls=art_b,w=1,s=nil,ew=0,lu=art_b,saw=&b}

{wnra=16,wnrr=12,snr=2,**ori=reforma**,pctr=no,last=no,pctl=no,offset=93,lw=no,gra=small,nb=sg,g=f,ds=reforma,ls=reforma,**c=noun**,w=1,s=nil,ew=0,lu=reforma,ts=reforma,t=reforma,error=46,ns=Reforma,ehead={case=nom;acc,nb=sg,g=f},saw=&b}

{wnra=16,wnrr=12,snr=2,**ori=reforma**,pctr=no,last=no,pctl=no,offset=93,lw=no,gra=small,**vtyp=fiv**,**tns=pres**,**mode=ind**,**per=3**,nb=sg,ds=reformar,**ls=reformar**,**c=verb**,w=1,s=nil,ew=0,lu=reformar,ts=reformar,t=reformar,saw=&b}

{wnra=17,wnrr=13,snr=2,**ori=del**,pctr=no,last=no,pctl=no,offset=97,lw=no,gra=small,**c=w**,**sc=p**,pcomp=yes,subl={reg=0,dom=0,loc=no,freq=0},ds=del,ls=del,w=1,s=nil,ew=0,lu=del,saw=&b}

{wnra=18,wnrr=14,snr=2,**ori=paro**,pctr=yes,last=no,pctl=no,offset=102,lw=yes,gra=small,nb=sg,g=m,ds=paro,ls=paro,**c=noun**,w=1,s=nil,ew=0,lu=paro,ts=paro,t=paro,error=46,ns=Paro,ehead={case=nom;acc,nb=sg,g=m}}

{wnra=18,wnrr=14,snr=2,**ori=paro**,pctr=yes,last=no,pctl=no,offset=102,lw=yes,gra=small,**vtyp=fiv**,**tns=pres**,**mode=ind**,**per=1**,nb=sg,ds=parar,**ls=parar**,**c=verb**,w=1,s=nil,ew=0,lu=parar,ts=parar,t=parar}

{wnra=18,wnrr=14,snr=2,**ori=paro**,pctr=yes,last=no,pctl=no,offset=102,lw=yes,gra=small,**vtyp=fiv**,**tns=pres**,**mode=ind**,**per=1**,nb=sg,ds=parir,**ls=parir**,**c=verb**,w=1,s=nil,ew=0,lu=parir,ts=parir,t=parir}

Each word shows all its possible interpretations and information. For example, *reforma* (reform) is analyzed firstly as a noun (*c=noun*), and secondly as the third person present indicative of the verb *reformar* (to reform) (*c=verb,ls=reformar,vtyp=fiv,tns=pres, mode=ind,per=3*).

From a linguistic point of view, words are far more interesting than terms, as the load of information words contain can be –depending on the type of word– syntactically transparent or ambiguous. For example and as shown above, words can have several interpretations and, therefore, belong to various syntactic classes and have alternative linguistic information. On the contrary, terms are input generally as invariable words in the dictionary of morphemes, that is to say, not as morphological roots but as full forms –like *aterosclerosis* in the following entry from the **dictionary of morphemes:**

{**string=aterosclerosis**,lu=aterosclerosis,c=n,t=nsing,g=f,s=disease}

In case MPRO cannot find any information in the lexica after analyzing a word or string of characters, the program assigns the word by default the syntactic category "noun" and marks it as "unknown". For example, in the following morphological analysis *valvulopatía* (valvulopathy) is not recognized by MPRO:

{**ori=valvulopatías**,saw=&r&n,wnra=12,wnrr=12,snr=1,gra=small,pctr=no,pctl=no,last=yes,**state=unknown**,**c=noun**,s=n,lu=valvulopatías,ds=valvulopatías,ls=valvulopatías,w=1,ehead={case=nom;acc,nb=sg;plu,g=m;f},cat=noun,case=nom;acc,nb=sg;plu,g=m;f}

3.2. *PARSER: syntactic analysis*

The module PARSER carries out the syntactic analysis and takes as input the results obtained from the morphological analysis done by LESEN. PARSER consists of a series of programs and a procedural explicit grammar (called **newgrammar)**, written

in an extremely simple formalism. This grammar is divided into four subgrammars that contain series of phrase structure rules –approximately 1200.

To the effect of the syntactic analysis of Spanish texts, PARSER applies the subgrammars sequentially; that is, it starts with the first grammar, then with the second and so through to the fourth. When all rules from the first subgrammar have been tried out and at least one of them works, the program applies once more the same subgrammar to ensure that only the rule selected is correct. If all rules from the first grammar fail to apply, the program jumps to the next subgrammar and tests new rules.

Likewise, the application of the rules is sequential because it follows the order in which they are distributed in the subgrammars. Finally, when all conditions of a rule are fulfilled, the program builds the syntactic structure described in the rule from the morphological analysis taken as input (Maas 1998: 168).

Since syntactic ambiguity cannot be solved once analysis is over –as there is no possibility of backtracking–, it is necessary to maintain an adequate sequence of the rules, which allows to cater for a variety of contexts of the structures described, with a view to minimizing the error margin (Maas 1998: 168).

After syntactic rules are applied, the program generates automatically a file with the results of the complete morphosyntactic analysis (*scr.ana*) shown as a tree graphic, where every element pertaining to a syntactic group of the sentence is assigned a feature bundle.

3.2.1. MPRO's Grammar

MPRO analyses sentences, namely it analyses the sentence structure and its constituents following the X-bar analysis. The parser starts from the simplest relations and projects them into more complex ones thanks to unification rules. That is to say, MPRO recognizes sentence constituents as noun phrases (np), verb phrases (vp), adjectival phrases (ap) and prepositional phrases (pp), etc., which are built from simple word categories like nouns, verbs, adjectives, etc. The X-bar sub-theory predicts the internal structure of the phrases stemming from the four basic lexical categories noun, verb, adjective and preposition. These basic categories, in turn, "project" their syntactic properties in the sentence formation process (Chomsky 1997: 19). For example, if we take an X category, its maximum projection will be PX (phrase X) and the PX's category will remain its nucleus' category, that is to say X. The four basic lexical categories also accept a "complement" to their right, like N" or P", which means noun+complement and verb+complement. As a result, basic lexical categories have more than one projection (SN, SV, SA, SP), as dictated by rules of phrase structures. This also means that phrases are nor flat nor lineal. Instead, they show several internal levels, like intermediate syntactic structures (N',V',A',P') between the item (N,V,A,P) and the traditional phrases.

Following the above philosophy, the grammar for syntactic analysis developed in MPRO is based on the phrase structure grammar. Hence, the starting point of the grammar is a formal and limited specification of series of parts of sentences that generate a set of possible sentences in Spanish. This is so since the construction of any sentence

results from the application of a specific set of rules from the grammatical system. This means that our starting point is not the sentence, but a part of it which, on the basis of predefined rules, aims to find an upper projection by relating itself to the rest of the elements of a string of signs on its left. In a nutshell, the rationale of our grammar specifically for Spanish has been to reconstruct syntactic relations from right to left, taking punctuation signs as discourse delimiters and starting points. If we take, for example, the sentence *"el perro es el mejor amigo del hombre, cuando no muerde" (the dog is man's best friend when it doesn't bite),* MPRO starts the projection from the entries – previously disambiguated by the morphological processor – to the phrases. The analysis begins by the first item in the string of signs which corresponds to a punctuation sign such as a period (.) or, in our example, a comma (,).

To show how a syntactic analysis of MPRO looks like, we present here the results of the morphosyntactic analysis of the following Spanish sentence taken from the newspaper article "Reforma del desempleo" appeared in the Spanish newspaper *El País* in September 2002: "*Aznar ofrece diálogo a los sindicatos pero elude hablar de la reforma del paro*" (Aznar offers the syndicates a dialogue but refuses to talk about the reform of the system of unemployment benefits):

<1> bis <15>: hs
{**cat=hs**,**r=328**,snr=2}
{**cat=hs**,**r=307b**,snr=2}
{**cat=hs**,**r=307b**,snr=2}
{**cat=hs**,**r=307a**,snr=2}

{**ori=Aznar**,saw=&b,wnra=5,**wnrr=1**,snr=2,gra=cap,pctr=no,pctl=yes,last=no,ew=1,lw=no,lu=Aznar,**c=noun**,g=m,nb=sg,s=surname,ds=Aznar,ls=Aznar,w=1,fw=yes,key=Aznar,ehead={case=nom;acc,nb=sg,g=m},cat=noun}

{**ori=ofrece**,saw=&b,wnra=6,**wnrr=2**,snr=2,gra=small,pctr=no,pctl=no,last=no,ew=0,lw=no,vtyp=fiv,tns=pres,mode=ind,per=3,nb=sg,lu=ofrecer,ds=ofrecer,ts=ofrecer,ls=ofrecer,t=ofrecer,**c=verb**,w=1,s=nil,**cat=fiv**}

{**ori=diálogo**,saw=&b,wnra=7,**wnrr=3**,snr=2,gra=small,pctr=no,pctl=no,last=no,ew=0,lw=no,nb=sg,lu=diálogo,g=m,s=text,ds=diálogo,ts=diálogo,ls=diálogo,t=diálogo,c=noun,w=1,ehead={case=nom;acc,nb=sg,g=m},**cat=noun**}

{**cat=pp**,**r=13a**,lu=a,snr=2}

{**ori=a**,saw=&b,wnra=8,**wnrr=4**,snr=2,gra=small,pctr=no,pctl=no,last=no,ew=0,lw=no,lu=a,**c=p**,pcomp=no,ds=a,ls=a,w=1,fw=yes,key=a,s=nil,cat=p}

{**cat=np**,**r=10,**fu=np,ehead={case=nom;acc,nb=plu,g=m},snr=2}

{**ori=los**,saw=&b,wnra=9,**wnrr=5**,snr=2,gra=small,pctr=no,pctl=no,last=no,ew=0,lw=no,lu=art_b,**c=w**,**sc=art**,nb=plu,g=m,fu=a,ds=art_b,ls=art_b,w=1,fw=yes,key=los,s=nil,cat=art,ehead={case=nom;acc,nb=plu,g=m}}

{**ori=sindicatos**,saw=&b,wnra=10,**wnrr=6**,snr=2,gra=small,pctr=no,pctl=no,last=no,ew=0,lw=no,nb=plu,lu=sindicato,g=m,ds=sindicato,ts=sindicato,ls=sindicato,t=sindicato,**c=noun**,w=1,s=nil,ehead={case=nom;acc,nb=plu,g=m},cat=noun}

{**ori=pero**,saw=&b,wnra=11,**wnrr=7**,snr=2,gra=small,pctr=no,pctl=no,last=no,ew=0,lw=no,lu=pero,**c=w,sc=coord**,ds=pero,ls=pero,w=1,fw=yes,key=pero,s=nil,cat=coord}

{**cat=hs**,**r=440**,snr=2}

{**ori=elude**,saw=&b,wnra=12,**wnrr=8**,snr=2,gra=small,pctr=no,pctl=no,last=no,ew=0,lw=no,vtyp=fiv,tns=pres,mode=ind,per=3,nb=sg,lu=eludir,ds=eludir,ts=eludir,ls=eludir,t=eludir,**c=verb**,w=1,s=nil,**cat=fiv**}

{**ori=hablar**,saw=&b,wnra=13,**wnrr=9**,snr=2,gra=small,pctr=no,pctl=no,last=no,ew=0,lw=no,vtyp=inf,lu=hablar,ds=hablar,ts=hablar,ls=hablar,t=hablar,**c=verb**,w=1,s=nil,**cat=inf**}

{**cat=pp**,**r=215**,lu=de,snr=2}

{**cat=pp**,**r=13a**,lu=de,snr=2}

{**ori=de,**saw=&b,wnra=14,**wnrr=10**,snr=2,gra=small,pctr=no,pctl=no,last=no,ew=0,lw=no,lu=de,**c=p**,pcomp=no,ds=de,ls=de,w=1,fw=yes,key=de,s=nil,cat=p}

{cat=np,r=10,fu=np,ehead={case=nom;acc,nb=sg,g=f},snr=2}

{**ori=la**,saw=&b,wnra=15,**wnrr=11**,snr=2,gra=small,pctr=no,pctl=no,last=no,ew=0,lw=no,lu=art_b,**c=w**,**sc=art**,nb=sg,g=f,fu=a,ds=art_b,ls=art_b,w=1,fw=yes,key=la,s=nil,cat=art,ehead={case=nom;acc,nb=sg,g=f}}

{**ori=reforma**,saw=&b,wnra=16,**wnrr=12**,snr=2,gra=small,pctr=no,pctl=no,last=no,ew=0,lw=no,nb=sg,lu=reforma,g=f,ds=reforma,ts=reforma,ls=reforma,t=reforma,**c=noun**,w=1,s=nil,ehead={case=nom;acc,nb=sg,g=f},cat=noun}

{**cat=pp**,**r=34**,lu=de,snr=2}

{**ori=del**,saw=&b,wnra=17,**wnrr=13**,snr=2,gra=small,pctr=no,pctl=no,last=no,ew=0,lw=no,lu=de,**c=p**,**pcomp=yes**,nb=sg,g=m,ds=de,ls=de,w=1,fw=yes,key=del,s=nil,cat=p}

{**ori=paro**,wnra=18,**wnrr=14**,snr=2,gra=small,pctr=yes,pctl=no,last=no,ew=0,lw=no,nb=sg,lu=paro,ds=paro,ts=paro,ls=paro,t=paro,w=1,s=nil,ehead={case=nom;acc,nb=sg,g=m;f},**cat=noun**}

<15> bis <16>: punct

{**lu=.**,**c=punct**,ds=.,ls=.,w=1,fw=yes,key=.,s=nil,cat=punct,ori=.,saw=&n&b&n,wnra=19,**wnrr=15**,snr=2,gra=other,pctr=no,pctl=no,last=yes,ew=0,lw=yes}

As explained above, MPRO builds phrase structures from the most basic categories –verbs, nouns, adjectives, etc.– and then tries to join them in a larger structure of the type

of a sentence. The very first information at the top of the analysis shown above is the scope of the analyzed structure (*<1> bis <15>: hs*). This means that from word 1 to word 15 MPRO built a larger Structure, namely a main sentence (*cat=hs*), which in this case is correctly analyzed. Once the program identifies the limit of the sentence, that is to say the dot (*lu=.,c=punct*), it starts to build a structure from that point to the beginning of the text. Since it finds a noun (*cat=noun,ori=paro*) and a preposition (*c=p,ori=del*), it applies one of its grammar rules (*r=34*) and builds a prepositional phrase (*cat=pp*). Again MPRO finds a similar structure: a noun (*c=noun,ori=reforma*), and article (*c=w,sc=art,ori=la*) and a complex preposition (*cat=p,ori=del, pcomp=yes*), which corresponds to one of its grammar rules on prepositional phrases (*r=13a*). Then, another rule for prepositional phrases is found (*cat=pp,r=215*), which enables to build a larger pp from the two already found. In this way, the MPRO-grammar builds the next phrase, which in this case is a complete sentence (*cat=hs*) composed of a finite verb form (*c=verb,cat=fiv,ori=elude*) and an infinitive verb form (*c=verb,cat=inf, ori=hablar*). Later on, a rule for nominal phrases is found (*cat=np,r=10*), which allows to join an article (*c=w,sc=art,ori=los*) with a noun (*c=noun,ori=sindicatos*) and a coordination (*c=w,sc=coord,ori=pero*). The final two phrases MPRO finds are, first, a pp (*cat=pp*) with rule 13a – and, second, a main sentence (*cat=hs*) with rule number 307a. This last main sentence is formed by a noun (*c=noun,ori= Aznar*), a finite verb (*c=verb,cat=fiv, ori=ofrece*) and another noun (*cat=noun ,ori=diálogo*). After that, the program finds the rule 307b for main sentences and applies it two times. The reason for the recurrent application of rule 307b, is that this rule enables to build a main sentence composed of another main sentence and a noun or a preposition, and this structure is found two times: the fist time with "*Aznar ofrece diálogo a*" and the second time with "*Aznar ofrece diálogo a los sindicatos*". Finally, MPRO finds rule number 328 and builds the whole sentence (*cat=hs,r=328*), which is composed of the main sentence (*cat=hs,r=307b*) –"*Aznar ofrece diálogo a los sindicatos*"–, a coordination –"pero"– and another main sentence (*cat=hs,r=440*) –"elude hablar de la reforma del paro".

4. Work to improve morphological analysis

Below are described some of the measures taken to improve the morphological analysis of Spanish texts in general with MPRO. The improvements reached so far cannot be regarded only from the number of words correctly analyzed. They should be also considered regarding the synthesis and consistency of the information coded in the lexica, which have been possible after the application of some strategies described in the following subsection.

4.1. *Verb morphology: inflection*

This section describes the strategy implemented to code verbs in MPRO in order to obtain correct morphological analyses of the existing verb forms in Spanish. To code

the verbs in the dictionaries, our bibliographical reference has been the *Langenscheidts Handwörterbuch (Spanisch-Deutsch)* (1998), which includes a two and three-fold classification of most Spanish verbs according to their conjugation: first, second and third. Our guide for verb lexemes has been Sánchez (1972).

In Spanish, irregular verb forms may present either changes in the roots or variations in their endings. In MPRO it is in the **file of inflected forms** where, through unification and feature sharing processes, verb roots and ending variations are dealt with. Such variations are intended, for example, to preserve phonemes – e.i. *remozar (to renovate)*, *remocé (I renovated)*, *vencer (to defeat)*, *venzo (I defeat)*; *afligir (to afflict)*, *aflijo (I afflict)*; *tocar (to touch)*, *toque (I touched)*, *rogar (to beg)*, *rogue (I begged)*; *averiguar (to find out)*, *averigüé (I found out)*.

To exemplify the coding strategy, we have chosen the verb *pensar (to think),* which has a double root (*pens; piens*) and is subject to several morphological changes, depending on some grammatical information such as tempus (tense), modus (mood), etc. – consider for example *pensó (he/she thought), pensé (I thought), pensábamos (we thought), pensaríamos (we would think)*. This is how the two roots of *pensar* are coded in the **dictionary of morphemes**:

{**string=pens**, c=v, **lu=pensar**,n={ador=agen,amiento=process,**t=ar-voc**}

{**string=piens**, c=v, **lu=pensar**,**t=piens**}

The value of the feature "t" refers to the verb flexion that the roots with such "t" value can have depending on person, tempus, modus, etc. In our example, the values of "t" are the verb class *ar-voc* for the root "pens" and the verb class *piens* for the root "piens". The inflections which correspond to these verb classes are coded in the special **file of inflected forms**. To illustrate how inflections are coded in the file of inflected forms, we present the following feature bundle which describes one of the possible inflections of *pensar* and of other verbs: the allomorpheme "o" which corresponds, among others, to the first person, indicative, present. The value of "string" is the inflection morpheme –here "o"– and the values of "l" are the verb classes *sient*, *piens*, *ar*, *fragu* and *1pps*:

{**string=o**,c=flex,**l=sient;piens;ar;fragu;1pps**,m={vtyp=fiv,tns=pres,mode=ind,per=1,nb=sg}}

In the first line, "*string=o*" is the corresponding inflection for present tense, indicative, first person and singular of verbs classes such as "sient", "piens", "ar", "fragu" and "1pps". With these unification rules, the system identifies and analyses e.g. "pienso" as the present tense, indicative, first person singular of *pensar*.

This codification synthesizes the analysis of verbs that share common features with *pensar* such as *contar (to count)*, which requires also two entries in the dictionary of morphemes that correspond to the verb's possible roots in all tenses and moods, that is to say *cont* and *cuent*:

{**string=cont**,**lu=contar**,c=v,**t=ar-voc**,n={ador=agent,a=able=able;ante=va}}

{**string=cuent**,**lu=contar**,c=v,**t=piens**}

Compound verbs such as *desmontar* (to dismantle) and *recontar* (to re-count), and others like *cerrar* (to close), *errar* (to miss), *augurar* (to prophesy) and *jugar* (to play) are coded after the same model of *pensar*.

Many of the verbs coded in the dictionaries are complex and have different roots. This has forced us to code each root ("string") and specify its corresponding endings ("t") according to tense, mood, etc. The example below shows the multiple roots in the verbal paradigm of the Spanish verb *tener* (to have):

{**string=ten**,lu=tener,c=v,**t=ten;dr**}

{**string=tien**,lu=tener,c=v,**t=tien**}

{**string=teng**,lu=lu=tener,c=v,**t=pressubj;1pps**}

{**string=tuv**,lu=tener,c=v,**t=pret**}

The codification of the roots becomes more complex when it comes to verbs belonging to the third conjugation. An example is the verb *seguir (to follow)*, that belongs to the third conjugation of verbs ending in *–ir*. Its codification is particularly complex due to the variety of roots that need to be coded for the system to be able to recognize them:

{**string=segu**,lu=seguir,c=v,n={imiento=process},a={ible=able},**t=ir**}

{**string=sig**,lu=seguir,c=v,**t=pressubj**}

{**string=sigu**,lu=seguir,c=v,**t=piens**}

{**string=sigui**,lu=seguir,c=v,**t=pret**,a={ente=va}}

{**string=siguier**,lu=seguir,c=v,**t=er-voc;impsubj1**}

4.2. *Derivational Phenomena: prefixation*

Another interesting result accomplished in MPRO-Spanish concerns the prefixation. To this end, we have established a division between "free" and "fixed" prefixes. Free prefixes are coded in the **file of inflected forms** and the program analyses them when it comes across words that have them. This implies for example, that in order for the program to analyse correctly "oligoelemento", there is no need to add the prefix "oligo" in the dictionary entrance for "elemento", as it was the case before. This is particulary helpful when working with specialized/medical/scientific texts which abound in prefixes such as "oligo", "piro", "sofo", "higro" or "picto". Otherwise it would be necessary to code each possible prefix on every dictionary

entrance which allows prefixation. The following is the dictionary entry of the word *elemento* (element) which does not have anymore the feature bundle "prf" for prefix. As a result, MPRO can analyse any of these free prefixes as such when combined with *elemento*:

{**string=element**,lu=elemento,c=n,t=o,g=m}

On the other hand, fixed prefixes need to be coded in the dictionary entries of those words that can combine with them. For example, fixed prefixes such as *ante*, *com*, *contra*, *de*, *descom* and *dis* are specified in the feature bundles of all roots of the verb *poner*. Some of the possible fixed prefixes that the roots of the verb *poner* (to put) can have are the following:

{**string=pon**,lu=poner,t=er,**prf=ante;com;contra;des;com;dis**}

{**string=pondr**,lu=poner,t=er,**prf=ante;com;contra;des;com;dis**}

{**string=pus**,lu=poner,t=er,**prf=ante;com;contra;des;com;dis**}

The fact that MPRO has been conceived as a system whose starting point is the lexeme rather than the word allows the dictionaries to be more synthetic and this explains why words such as *anteponer* (to place before), *deponer* (to lay down), *descomponer* (to break down) or *disponer* (to lay out) are not analysed as lexemes but rather as prefixed words of the lexeme *poner*. The following are the results of the analysis of MPRO for the words mentioned above:

{wnra=1,wnrr=1,snr=1,**ori=anteponer**,pctr=yes,last=no,pctl=no,offset=10,lw=no,gra=small,vtyp=inf,ds=ante$poner,ls=ante$poner,c=verb,w=1,s=nil,ew=1,lu=anteponer,ts=anteponer,t=infinitiv }

{wnra=7,wnrr=7,snr=1,**ori=deponer**,pctr=yes,last=no,pctl=yes,offset=42,lw=no,gra=small,vtyp=inf,ds=de$poner,ls=de$poner,c=verb,w=1,s=nil,ew=0,lu=deponer,ts=deponer,t=infinitiv}

{wnra=9,wnrr=9,snr=1,**ori=descomponer**,pctr=yes,last=no,pctl=yes,offset=55,lw=no,gra=small,vtyp=inf,ds=des$componer,ls=des$componer,c=verb,w=1,s=nil,ew=0,lu=descomponer,ts=descomponer,t=infinitiv}

{wnra=11,wnrr=11,snr=1,**ori=disponer**,pctr=yes,last=no,pctl=yes,offset=65,lw=no,gra=small,vtyp=inf,ds=dis$poner,ls=dis$poner,c=verb,w=1,s=nil,ew=0,lu=disponer,ts=disponer,t=infinitiv}

{wnra=13,wnrr=13,snr=1,**ori=poner**,pctr=yes,last=no,pctl=yes,offset=72,lw=yes,gra=small,vtyp=inf,ds=poner,ls=poner,c=verb,w=1,s=nil,ew=0,lu=poner,ts=poner,t=infinitiv}

5. Some experiments with MPRO

This section is devoted to present the results of some experiments realized with MPRO –like *term extraction*, *indexation* and *information retrieval*– in order to test some of the program's applications.

5.1. *Term extraction*

Terminological Extraction (TE) is one of the applications in the area of natural language processing that is currently being implemented with MPRO. With base on a correct morphological analysis and a program to extract terms it is possible to extract terminology from a text. The reasons for using a morphosyntactic parser like MPRO for automatic term extraction are twofold: first, its capacity to recognize noun phrases with a minute error margin. Secondly, it helps identify noun phrases that are treated as terms in a specific field, which must be previously specified in the dictionary (Hong et al. 2001: 4). The program used for term extraction is called *AutoTerm* and, like MPRO, was developed at the IAI.

AutoTerm classification parameters for term candidates are based not only on morfosyntactic information but also on statistics. A noun or noun phrase is considered to be a term candidate on the basis of: a) its frequency of appearance in the text (in other words, based on its "importance"), and b) accordingly to the number of times its components –i.e. two nouns or a noun and an adjective– appear associated in the same syntactic phrase (comp. Hong et al. 2001: 6).

A test for terminological extraction for Spanish was carried out using a text on electronics of approximately 4,000 words. As many as 359 words were considered term candidates. The following table shows some of the results of the extraction. It contains 15 term candidates with their corresponding "weights". The first term candidates on the list have higher values or "weights" than the last ones. This in turn means, that the first term candidates are more likely to be terms – i.e. *sistema de información* (information system) –, whereas the last ones on the list – i.e. *caso de diseño* (design case) – are less likely to be considered terms:

70. sistema de información	345.17
87. entorno tecnológico	125.32
103. soporte técnico	106.47
...	
168. estándar técnico y de nomenclatura	11.29
169. elemento afectado	11.29
170. directriz tecnológica o de integración	11.29
171. determinada solución de infraestructura	11.29
...	

355. parte del sistema	2.13
357. especificación de excepción	1.99
359. caso de diseño	1.47

Among these terms, a minute amount of them is composed of just one noun like *entrada* (entry, entrance) and *salida* (exit). Most term candidates extracted by the program, as many as 80% of them, were composed of a noun + adjective, a noun + noun phrase or a noun + prepositional phrase. According to the results obtained, we can conclude that around 60% of the candidates are made up of at least one term.

In order to reduce the number of false term candidates AutoTerm is fitted with a list of restrictors, that is, a list of expressions and words –i.e. adjectives and adverbs– that do not have anything to do with termhood and consequently cannot be parts of terms. Among others, some of the expressions considered restrictors for Spanish are the following: *a causa de* (because of), *a condición de que* (on condition that), *a continuación* (next), *a fin de que* (so that), *a la vez que* (as), *a medida que* (as long as), *a menos que* (unless), *a partir de* (from), *a pesar de* (in spite of), *con respecto a* (as regards), *con tal de que* (as far as), *conforme a* (according to), *contra* (against), *cual* (which), *cualquier* (any), *dicho* (abovementioned), *en absoluto* (at all), *en calidad de* (as a/an), *en caso de que* (in case that), *en cuanto a* (as regards), *en el caso de que* (in case that), *en la suposición de que* (supposing that), *en tanto que* (as far as/as long as), *enfrente de* (in front of), *entre* (between), *entretanto* (meanwhile), *mas* (but), *más de* (more than), etc.

5.2. *Indexation*

Text description and indexing are further applications of MPRO, for which semantic information is needed. This is currently being introduced in MPRO-Spanish. Until now, indexing was only possible for English and German texts with the indexing program BINDEX, which has been developed also at the IAI. Now Spanish is also available to the program. In order to index Spanish texts, some of the modules of BINDEX were first adapted to take the Spanish morphosyntactical analysis as input and a Spanish thesaurus have been created on the basis of translations from existing German and English thesauri. The Spanish thesaurus, which contains 49.685 entries, is now being revised and "cleaned". The process of text description and indexing uses statistical techniques which allow identify semantic characteristics –along with their corresponding lexemes– that appear most frequently in the text. The indexing is calculated additionally from the semantic information of sentences thanks to semantic descriptors. At the end of the process, a maximum of ten sentences is generated, one of them being often the title of the text (Maas 1998: 171). For example, the following are some of the semantic descriptors used by BINDEX to describe a Spanish text about commercialization of graphic products in internet. The values in square brackets correspond to the frequency of appearance of the terms in the text and, therefore, are a sign of their potential to represent the meaning of the text.

Descriptors: comercio[100] (business); punto[85] (point); línea[68] (line); comando[61] (command); tipo[61] (type); conexión[48] (connection); relación[40] (relationship); red[35] (network); error[33] (error); plataforma[31] (platform); prueba[30] (test); inicio[30] (beginning); área[30] (area); caso[21] (case); superficie[20] (surface); programa[20] (program); razón[20] (reason); tarifa[20] (tariff); servicio[18] (service); centro[13] (centre); acción[13] (actino); llamada[12] (call); segmento[12] (segment); cliente[11] (client); método[11] (method).

The results obtained so far from the indexation of Spanish texts were revised manually and are considered to be satisfactory, because the semantic descriptors chosen by BINDEX correspond to the topic of the analysed texts. Nevertheless it is important to remark that these are preliminary results and that there still is work to be done regarding the revision of the thesaurus and the descriptors. At the same time, further tests are needed in order to determine the reliability of BINDEX with the Spanish morphosyntactical analysis of MPRO as basis.

5.3. *Information retrieval*

The linguistic analysis carried out by MPRO can also be used as input for knowledge-based information-retrieval systems. LEWI is an example of a lexicon-based information-retrieval system, which is now being developed at the IAI. The aim of the system is to give information contained in the German encyclopaedia *Brockhaus* by a question-answering process. In order to achieve this, the system must represent questions logically by using a thesaurus and semantic descriptors, which are also needed for indexing. Up to date, the system is intended for German only. Though, it has been possible to adapt LEWI's analysis modules, so that the Spanish thesaurus and lexica are used – instead of the German ones – and results of the linguistic analysis of Spanish texts are taken as input for the system. As a result of this first experiment for Spanish, the system has successfully identified Spanish question phrases and had delivered their logical representation. Though, it failed to give the semantic information needed for the interpretation of the logical representation. One of the reasons for this failure is that the Spanish thesaurus still needs to be checked and more descriptors must be added, both processes which represent future work. As an example, the logical representation of the Spanish sentence *¿Cuándo estalló la guerra civil española?* (When did the Spanish Civil War break out?) is shown next:

estallar[] (({guerra,civil,español}),()).

Judgment question

The system extracts the verb *estallar (break out)* as the nucleus of the question, it identifies the subject –in brackets – and recognises the question as a "judgment question". So far this is the only type of question that the system recognises for Spanish.

The results of this experiment are preliminary. We intended to test if it was possible to work with LEWI based on the morphological analysis obtained with MPRO for information retrieval. During the experiment, we did not test any other program for information-retrieval. This can in turn be included now among the tasks of a future work. Another future line of work is the classification of questions for Spanish and its integration in LEWI.

6. Summary and future research

The Spanish parser MPRO presented in this article is a program for morphosyntactic analysis of general and specialized texts which, at the moment, is still at an experimental level, albeit the progress described, i.e. the work done regarding verbs, prefixes, the revision of entries in the dictionaries, the addition of new entries as well as the elaboration of new syntactic rules for noun phrases in the grammar, is encouraging. Broadly speaking, we have already managed to synthesize the codification strings in such a way that one string contains the morphological, syntactic and semantic information of the word and all its possible derived forms: for example, adjectives contain the information of possible derived adverbs and verbs are coded also with information about possible derived nominalizations or adjectives.

There are several lines of research that are already under way. Two applications of the parser, indexing and terminological extraction are currently underway. These applications are forcing us to revise thoroughly all dictionaries, syntactic rules for noun phrases in the grammar and check the entries of the Spanish thesaurus.

Further future work will focus on intensive effort towards disambiguation both at morphological and syntactic level as well as semantic. Semantic information is becoming increasingly necessary to analyze sentences whose meaning cannot be disambiguated at syntactic level. Although semantic information has not been used systematically in the codification of the program, some of it has already been introduced in LESEN. For future and more systematic work on semantics we have already looked at the possibility of using lexical databases[5]. Another aspect that needs further work is the treatment of "unknown words". For the time being, the program cannot assign the unknown word a category other than noun. One of our aims is to get the program to classify unknown words not only as nouns, but also as verbs and adjectives, depending on what their categories are.

Notes

1. The IAI is the Institute of Applied Information Sciences of the University of Saarland (Germany).
2. El IAI es el instituto de Ciencias Aplicadas de la Información de la Universidad de Saarland.(Alemania).
3. The *tagging* of a text is defined as a series of linguistic processes which allow to *annotate* –or to enrich a text or corpus with linguistic information (syntactical information such as type of sentence, subject, object)– and recognize sentence structures (comp. to Zilinsky 2002: 27). *Tagging* and *parsing* will be treated as synonyms in the present article (comp. to Zilinsky 2002: 28).

4. MPRO's architecture is undergoing now a series of changes which aim at integrating all languages in a unique platform and with a user-friendly interface.
5. Although we are aware of the existence of semantic networks such as EDR and Roget's Thesaurus, we are envisaging using WordNet because of its extensive range and frequent use in other text-classification tasks such as *information retrieval* and *disambiguation* itself.

REFERENCES

Aone, C. and K. Hausman. 1996. "Unsupervised Learning of a Rule-based Spanish Part of Speech Tagger". *Coling-96*, *Proceedings*, Vol. 1, August 5-9, Denmark.

Badia, T., Egea, A. and A. Tuells. 1997a. CATMORF: "Multi two-level steps for Catalan morphology". *Technical Report*. [Available at: acl.ldc.upenn.edu/A/A97/A97-2015.pdf].

Badia, T., Egea, A. and A. Tuells. 1997b. CATMORF: "Multi two-level steps for Catalan morphology*". Demo Proceedings of ANLP97*.

Badia, T., Pujol, M., Tuells, A., Vivaldi, J., de Yzaguirre L. and T. Cabré. 1998. "IULA's LSP Multilingual Corpus: compilation and processing". Paper presented at the *ELRA conference*, Granada, 29-31 May 1998.

Berschin, H., Fernández-Sevilla, J. and J. Felixberger. 1995. "Die Spanische Sprache – Verbreitung, Geschichte, Struktur. –2.", aktualisierte Aufl. – Ismaning: Hueber, 1995.

Chanod, J-P., Hobbs, J., Hovy, E., Jelinek, F. and M. Rajman. 2001."Methods and Techniques of Processing". *Multilingual Information Management: Current Levels and Future Abilities*. Linguistica Computazionale. Volume XIV-XV. Eds. E. Hovy, N. Ide, R. Frederking, J. Mariani, and A. Zampolli. Pisa: Instituti Editoriali e Poligrafici Internazionali. [Also available at: http://www-2.cs.cmu.edu/~ref/mlim/chapter6.html]

Annotated corpus (CREA) [Available at: http://www.rae.es/rae/gestores/gespub00…68C1256ADB003663B7/$FILE/Anotacion2.htm] Corpus de Referencia del Español Actual (CREA). [www.rae.es]

Diccionario General llustrado de la Lengua Española (1987). Barcelona: Bibliograf, S.A.

Di Sciullo, A. and F. Sandiway. "Efficient Parsing for Word Structure" [Available at: http://www.afnlp.org/nlprs2001/pdf/0034-03.pdf.]

González, J. 1995. "ARIES: a ready for use platform for engineering Spanish-processing tools". *Language Engineering Convention*, London, October 1995. [Available at: [http://www.mat.upm.es/~aries/docs/lec95.ps. Stand: 08.03.03]

Hernanz, M.L. and J.M. Brucart. 1987. *La sintaxis, 1. Principios teóricos. La oración simple*. Barcelona: Editorial Crítica.

Hong, M., Fissaha, S. and J. Haller. 2001. "Hybrid Filtering for Extraction of Term Candidates from German Technical Texts". *Proceedings of the Conference TIA-2001*, Nancy, 3 et 4 mai 2001.

Langenscheidts Handwörterbuch (Spanisch-Deutsch) (1998): Berlin and München.

Maas, H-D. 1998. "Multilinguale Textverarbeitung mit MPRO". Lobing, G. (Hgs.), *Europäische Kommunikationskybernetik heute und morgen*. München.

Marín, F.M. and P. E. Ramírez. 2001. *Guía de gramática de la lengua española*. Madrid: Espasa Calpe.

Moliner, M. 1990. *Diccionario de Uso del Español* (2 volumes). Madrid: Gredos.

Rojo, G. 2001. "La explotación de la Base de datos sintácticos del español actual (BDS)". *Corpus lingüísticos*. Ed. J. Kock. Universidad de Salamanca.

Ruiz, L. 2000. "Etiquetación automática en corpus textuales cubanos. Primeros resultados». *JADT 2000 5° Journées Internationale d'Analyse Statistique des Donées Textuelles*. [Available at: http://www.cavi.univparis3.fr/lexicometrica/jadt/jadt2000/pdf/68/68.pdf]

Sánchez, M. 1972. *Prontuario de Conjugación. Diez mil verbos castellanos*. París: Bouret.

Subirats, C. 1998."Automatic Extraction of Textual Information in Spanish". *Language Design. Journal of Theoretical and Experimental Linguistics* 1: 1-13. [Available at: http://Seneca.uab.es/csubirats/Automatic.doc]

Subirats, C. and M. Ortega. 2002 "Extracción de información de grandes hábeas". *La lingüística de corpus: aplicaciones*. Eds. J. De Kock and C. Gómez. Salamanca: Ediciones Universidad de Salamanca. [Available at: http://Seneca.uab.es/csubirats/Automatic.doc]

Zielinski, D. 2002. *Computergestützte Termextraktion aus technischen Texten (Italienisch)*. Final Work submitted for a diploma in Translation at the University of Saarland.