

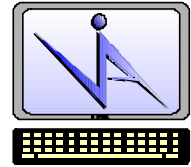
UNL Project 1ST Year Final Report

Universidad Politécnica de Madrid (UPM)
Madrid – Spain



Grupo de Validación y aplicaciones Industriales

Departamento de Inteligencia Artificial
Facultad de Informática
Universidad Politécnica de Madrid



1ST YEAR FINAL REPORT

Autors	:	Jesús Cardeñosa, Francisco Astudillo, David Escorial, Luis Iraola, Fermín Moscoso, Christèle Legeard
Institution	:	UPM
Project	:	UNL
Type	:	Deliverable
Date	:	19 th January 1997
Ref.	:	[UPM/UNL/DeI/V1.0/91]

Summary

1st Year Final Report and Annexes to be delivered to IAS describing the work carried out.

INTRODUCTION

This document describes the works carried out by the team of the Universidad Politécnica de Madrid (UPM) from December 1996 to November 1997 both included. This document has two parts. One is the description of the project and the achievements. The second is formed by the Annexes, which are the information referred to in this document and which are indispensable to understand the work done.

We have to mention explicitly all the people who participated to these works and who are, beside myself, the following persons:

Francisco Astudillo Pacheco (Doctor in Computer Science) who is living in Japan, contributed to this project by his work and his knowledge of great worth, **José Antonio Espinosa** (Computer Science Engineer) who, despite his young age, brought an experience of several years in subjects like that of this project, **Christèle Legard** (Trilingual Translator and Linguist) whose contribution proved to be essential for our working group, **Luis Iraola** (Doctor in Philosophy) who brought strength to our working group thanks to his experience in the Linguistic and Computer Science fields, **Viktoría Nefedova** (Computer Science Engineer) who contributed to the coding of the Dictionary, **David Escorial Rico** (Computer Science Engineer) who supported the main load of the generation task as his colleague is living abroad, demonstrating then that Internet is not only a network but also a means to work, **Monica Adanez Barba** (Computer Science Engineer) who contributed in great part to the writing of the Dictionary in the second phase of the project and when minutes seem to be seconds, **María del Carmen Carrasco** (Computer Science Engineer) who participated in the generator tasks in the second phase of the project, **Fermín Moscoso del Prado**, student of the last course of Computer Science Engineering and who is showing promise of having great capacities for the Natural Language field and, **Carlos Juárez Núñez**¹ (Computer Science Engineer) who was in charge of the document management in this project.

We must not forget to mention the **Fundación General de la UPM** which grants the institutional support to this project and which supported it efficiently at any time.

Jesús Cardeñosa
Director of the Spanish UNL Project

January 1998

¹ Guest Professor from the University of Atacama (Chile)

INDEX

1. GENESIS OF THE PROJECT	5
2. GENERAL ORGANIZATION OF THE PROJECT IN UPM	6
3. MANAGEMENT	8
4. THE ENCONVERTER (ANALYZER).....	11
4.1 Expected Results.....	11
4.2 Development of the works.....	12
5. DICTIONARY	21
5.1 Development of the Dictionary	21
6. DECONVERTER.....	31
6.1 Study of the Deco (software delivered by the IAS/UNU).....	31
6.2 Design of the rules:.....	31
6.3 Semantic Module.....	33
6.4 Syntactic Module.....	34
6.5 Morphologic Module.....	35
6.6 Implementation.....	36
6.7 Rule Debugging:.....	37
6.8 Work Environment	38
6.9 Glossary	38
6.10 Future And Conclusions	39
7. GENERAL REFERENCES	41
ANNEX 1 "Produced and Received Documents"	43
ANNEX 2 "Monthly Reports"	48
ANNEX 3 "Analysis Rules"	61
ANNEX 4 "Dictionary Contents"	94
ANNEX 5 "Generation Rules"	96
ANNEX 6 "Generation Results"	158

1. GENESIS OF THE PROJECT

In this section, we will report the succession of steps that made possible that two entities (UNU-IAS and UPM), without previous contacts between their Research Groups, started to work together in the Universal Networking Language Project (UNL).

At the early beginning of 1996, Mr. Francisco Astudillo Pacheco, employee of the Spanish Consulate in Tokyo and former collaborator of UPM Research Group in Artificial Intelligence subjects, contacted some persons related to the Institute of Advanced Studies (IAS) with a view to establish a collaboration between this Institute and his former Research Group.

Following these contacts, a direct relationship was established between Mr. Hiroshi Uchida and Mr. Sadaharu Takai (Director and Collaborator of the UNL Project respectively) and the Head of UPM Research Group, Mr. Jesús Cardeñoso, with a view to UPM participation in UNL Project in representation of the Spanish Language. After these previous contacts, IAS sent a skeleton of 3-year planning of this project for UPM team to study it and propose a work planning afterwards.

The project undoubtedly interested UPM as it has been a line of work for this Research Group since 1987, with its active participation in different international projects related to Linguistic Engineering. Among them, we can point out the participation of three of this Research Group current members in the PIVOT Project sponsored by NEC (Multimedia Laboratories, Tokyo). This team's experience in that project took place between 1987 and 1992 when NEC gave up the works. Between 1994 and 1996, the Group participated as a partner in two more projects on this field. One was the project called "PASO PC-315", "Natural Language Data Base Interface Generator", sponsored by the ESPRIT Program of the European Union and the Spanish Ministry of Industry. In parallel to this project, the team was a partner and distinguished participant of the ESPRIT II Project nº 8749, "ORCHESTRA", concretely in a Workpackage dedicated in part to the semantic analysis of documents.

After the reception of the provisional Planning for the UNL Project from the IAS, UPM team prepared a concrete proposal of tasks as well as the workload associated to these tasks. This document was the basis for the conversations between IAS and UPM maintained in July 1996 by Mr. Cardeñoso for UPM and Mr. Uchida as UNL Project Director for IAS. Mr. Takai, as a collaborator of Mr. Uchida, also came to this meeting. After the conversations, Mr. Uchida visited the Working Group installations, and then UPM just waited for IAS counter-proposal.

At the beginning of Autumn 1996, UPM received an e-mail in which IAS made a specific proposal of tasks to carry out by UPM and a global valuation of 6.500 h/m for the first year of the contract. Concretely, the tasks were:

1. Dictionary with 100,000 entries in Spanish language,
2. Linkage between dictionary and the Universal Words,
3. Morpho-syntactic grammar rules for analysis and generation between the Universal Networking Language and the Spanish Language
4. Submission of a brief monthly progress report to the United Nations University, with the corresponding work loads as indicated below.

Task Identification	Allocated Resources (m.h)
1.- Dictionary with 100.000 entries in Spanish Language	3.500
2.- Morphosyntactic Grammar Rules for Generator	2.000
3.- Morphosyntactic Grammar Rules for Analyzer	1.000

Table 1: Tasks and Resources for the first year

These were the essential conditions of the contract between both institutions. After this acceptance of the conditions, IAS convoked an inaugural meeting of all the Working Groups representatives from all the different countries participating in the Project. This meeting was to be held in the central headquarters of the IAS in Tokyo from 20th to 22nd November 1996. The essential representatives of UPM Working Group went to this meeting, specifically, José Antonio Espinosa, Christèle Legeard and Jesús Cardeñosa. Indeed, José Antonio Espinosa, Engineer in Computer Science and with experience in the Linguistics Engineering field from several years ago, Christèle Legeard, as a Linguist and multilingual translator and with experience in some of the projects mentioned before and Jesús Cardeñosa with a long experience in the Management and Development of international projects of this field, constitute the initial nucleus of UPM team to which the contribution of Francisco Astudillo Pacheco is added as an expert in the specific field of this project. It is an surprisingly experienced team in spite of its youth and with a high level of motivation for this project.

During Tokyo meeting, UPM contributed to the Workshop with some suggestions and its presentation and finally it determined with the persons in charge of the project the initial date for the beginning of the Project as 1st December 1996.

2. GENERAL ORGANIZATION OF THE PROJECT IN UPM

As any other project of these characteristics, its starting is essentially based in the work plannings and the assignation of resources that permit to carry these project out. As a result, the first month was dedicated essentially to these two matters: on the one hand, the configuration of the working group and on the other the elaboration and writing of the Internal Planning.

For the configuration of the working group and before the special characteristics of this project, we chose a model of increasingly temporal human resources assignation, that is, the working group is more reduced at the beginning and increases during the project to reach its maximum a the end of the first year. Likewise, the set of computing resources supported by the working group as well as those provided by the University were defined and planned. On the other hand, the first version of the Internal Work Planning was redacted. This plan broke down the global tasks of the contract into a group of Workpackages with tasks and subtasks in which, not only the tasks to be carry out, but also the period during which they had to be carry out, the person in charge of each task, etc., were defined. The scheme of typical task would be:

Task: <Task name>
Description: <Brief explanation of the work to do>
Person in charge: <Name of the person in charge>
Period: <Starting Date – Ending Date>
Resources: <Man/hour>
Input: <Documents or Software, or both (external or output of other tasks)>
Output: <Documents or Software, or both>

This Internal Planning was fitted into strict periods so as to assure the fulfillment of the External Planning that you can see below:

TASKS/SUBTASKS	m.h.	DC	JN	FB	MR	AP	MY	JN	JL	AG	SP	OC	NV
T.1. Generator of Spanish	2000	[Blue bar]											
<i>T.1.1 Previous Studies and Management</i>	200	[Grey bar]											
<i>T.1.2 Study of the Generator</i>	100	[Grey bar]											
<i>T.1.3 Rules Design</i>	1200	[Grey bar]											
<i>T.1.4. Rules Debugging</i>	300	[Grey bar]											
<i>T.1.5 Generation Tests</i>	200	[Grey bar]											
T.2 Dictionary	3500	[Blue bar]											
<i>T.2.1. Study of the Dictionary Format</i>	400	[Grey bar]											
<i>T.2.2 Selection of entries</i>	400	[Grey bar]											
<i>T.2.3. Writing of the Dictionary</i>	2200	[Grey bar]											
<i>T.2.4. Corpus Tests</i>	500	[Grey bar]											
T.3. Analyzer	1000	[Blue bar]											
<i>T.3.1 Previous studies</i>	100	[Grey bar]											
<i>T.3.2 Design of the Analyzer Rules</i>	600	[Grey bar]											
<i>T.3.3 Debugging</i>	200	[Grey bar]											
<i>T3.4 Tests</i>	100	[Grey bar]											
TOTAL	6500												

Table 2: Detailed Tasks, Resources and Duration

In this Internal Planning, tasks of all the kinds are gathered: those of training, management, technical, documentation, relations with the outside, considering the aspects of standards as well of organization of security and confidentiality.

The general organization of the project can be seen schematically in the following blocks diagram. (Figure 1)

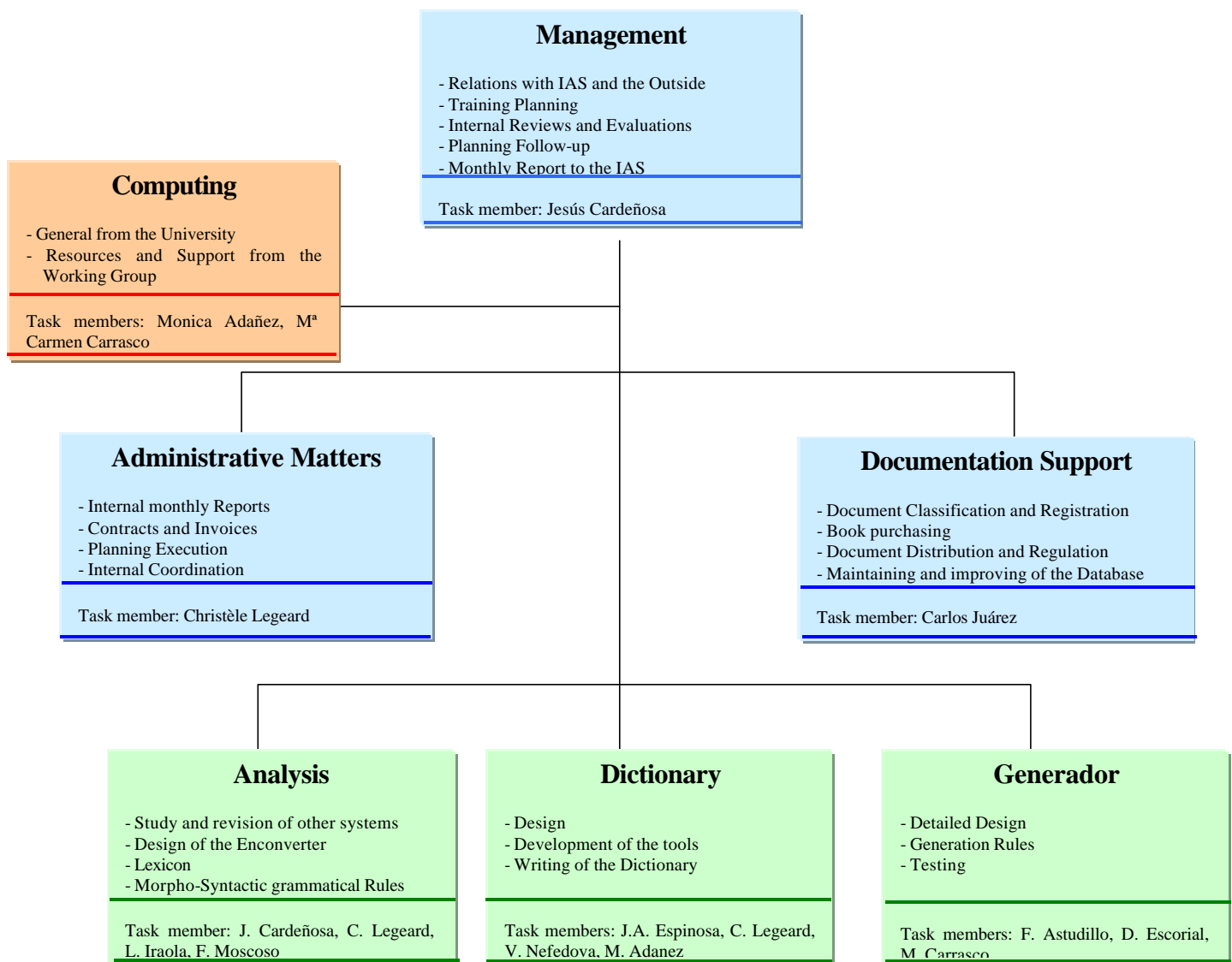


Figure 1: General Scheme of the UNL Project Organization

These blocks support the development of the project. Some of them is specific of this project, whereas others are already existing in the Research Group and are also supporting other projects. Nevertheless, even the already existing ones had to define precisely their contribution to the UNL project. We are going to describe each one of the tasks and functions defined for them.

3. MANAGEMENT

This task was in charge of Jesus Cardeñosa who in addition to the relations with the Director of the Project, Mr. Uchida, and other estates of the IAS, was in charge of defining and supporting all the **questions related to contracts and other legal questions**. In this task, he was helped by Christèle Legeard. In the beginning of the project, this task assumed also the responsibility to design an **internal training course** at the organizational level as well as the elaboration of the program. This course lasted two weeks and had a double direction. On the one hand, reviewing and training persons who would be part of the team henceforth in all the questions related to Linguistic

Engineering, that is, to train new persons who will be member of the team afterwards as well as to study thoroughly the initial material we were given during Tokyo Meeting.

This intensive course was imparted by Mr. Cardeñosa, Mr. Astudillo, Miss Christèle Legéard and Mr. José Antonio Espinosa, and were attended, in addition to themselves, by Miss Viktoria Nefedova, Mr. Fermín Moscoso, and Mr. Guillermo Pastor, who were from then on and during some time members of our team.

Part of the material used in the course was gathered together into a document which can be used as a basis for new possible incorporations. However this course demonstrated another important thing, it was that after it and its results, it was proved that someone with a superior formation in Computer Science could integrate our working group after a training period no longer than one month, even if he/she has no previous experience in Linguistic Engineering. This is obviously the case when the persons in charge of the different tasks do have this experience.

The internal following-up of the works was in charge of this task and an internal Planning that developed in details the tasks of the External Planning was designed for that purpose. This Internal Planning went through various modifications all over this year, due to the different external changes, mainly the delays in the reception of the Corpus from the IAS, as well as to the changes in UNL specifications. The following-up of this planning was made on the basis of two essential activities. One was the periodic global as well as task to task meetings of Jesus Cardeñosa with the members of his working group, and the other one was the monthly reports that all the members of the working group had to hand over to Jesús Cardeñosa and that were essential for the writing of the external monthly reports.

The other essential tasks related to this position are obvious, there are those of following-up and coordination of the other organizational blocks of the project, but particularly, those that are supporting the others, like the administrative tasks and the **Documentation and Archive Service** of the project. Among these tasks, we can point out:

- Definition of a planning that assures the confidentiality of the works, handling and access to the documentation generated.
- Definition of the format of all the documents to be generated in the project as well as their location form.
- Definition of a document Database of the project and of the archiving and registering processes.
- Service of localization of external documents, and research and purchasing of bibliographic material.
- Definition of processes for accessing the documentation generated.
- Maintenance and improvement in the Database adapting it to this project characteristics.

In ANNEX 1, you can see a list of the documents produced, received and in any case, registered, with an explanation on the registering formats and references.

Management was also in charge of guaranteeing at any time the **computing support** for all the activities of the project, as a result, the general resources provided by the University as well as those of the working group itself were defined.

Management was specifically in charge of the coordination and **production of all the external documents**, that, besides the Monthly Reports (ANNEX 2) were:

- Document on the State of the Project previous to Pisa Workshop [19]
- Document of support to the UPM Works Presentation to Mr. Uchida during its visit in Madrid on 16 October 1997. [48]
- Writing of this document. [60]

Management was also in charge of the **organization of the presentations and attendance to general meetings** held during this year. There were Pisa Workshop (Italy, 8-10 May 1997) and Paris Symposium held in UNESCO Headquarters (19-22 November 1997), where the works carried out during almost all the year were presented.

The last important attribution of this task was that of **establishing external contacts** in two directions. All of them were preliminary contacts but had a double orientation. The more notable was perhaps the one we have maintained with the Cervantes Institute, highest Spanish state institution for the Spanish Language promotion in the world. A meeting was held with the Academic Sub-Director, Joaquim Llisterri, to whom the project dissemination booklet was handed over, telling him besides the objectives of this project as well as the countries involved. This institution offered its collaboration in the evaluation of the works which, as they are not directly integrated in the project, could prove to be useful, that is, derived products. For example, one which was evaluated as interesting was a modification of the Enconverter or computer-assisted Analyzer to help the teaching of Spanish in the centers of the Cervantes Institute. The other useful point of this contact with this Institute for the Spanish part of the UNL was that we obtained precise information on the linguistic resources existing in Spanish and that could be used in the UNL. Unfortunately, we checked out that the resources existing in Spanish, and exploitable at the industrial level, were rather short.

Thanks to the participation of this Research Group in industrial projects in the Linguistic Engineering field, we maintained contacts with our former industrial partners so as to verify if there were whether resources or orientations that could permit the product or sub-products of the UNL to have uses in a shorter term than that estimated for the project itself. We did not find much. On the contrary, we verified what we already knew, that is, that in Spanish there are not any big linguistic resources that can have an industrial application at the computational level.

On the following we are going to describe the technical works carried out during this first period of the project. The basic scheme will follow the description of the initial planning and the expected results, the development of the works and finally the results obtained that will be described in these headings but that will also be supported by the necessary annexes at the end of this document. We separate the works reflected in the contract in three big blocks corresponding to the Analyzer or Enconverter, the Dictionary and the Generator or Deconverter.

4. THE ENCONVERTER (ANALYZER)

The agreement reached for this task for the first year as it reflected in the contract says that “*Morpho-syntactical rules for the analysis of the Spanish language*” will have to be defined or developed. Indeed any analyzer has to start by gathering a core that permits to determine the basic rules that represent its grammar. This must be, from the beginning, the previous step to the study and representation of semantics.

However, representing a grammar is not an easy task, even less with a view to a computational utilization. We must not forget that if today the efficiency of the natural language computational applications is rather short considering the resources invested, it is partly due to the fact that, especially in machine translation, the result researched was an automatic analysis. The deep fusion between morpho-syntax and semantics added to language dynamics has kept coming right up against this obstacle again and again. Hence, and following IAS indications, we have assumed that the Enconverter had to be assisted. It is important not to forget the goal of the Enconverter which is to assure that the texts in each language can generate a perfect translation to UNL. If not, the Generators can be good, but the generation will be wrong. So we think that the Enconverter is a critical point for the success of the project. As it is a critical point and as we assume that a complete automation of the linguistic analysis of every language is an impossible task today, we decided, as we said before, that the Enconverter would be assisted. By assisted we mean that the analysis tool defines a type of user who would be a linguist and who would translate to UNL with the help of the Enconverter. To define which part will depend on the help and which will depend on the system itself will be one of the goals of the second year, as it is indispensable to implement the first version of such a system.

So for the first year tasks, the emphasis was put on the definition of a core of morpho-syntactic rules the more general as possible so as to constitute a first core that could be automated successfully.

4.1 Expected Results

Considering the fact that the Analyzer has to be assisted, this will be the dominant element in all the actions of the working group from the beginning. Then the first year of work had to focus on various questions that can be summed up as follows:

4.1.1 Determination of the grammatical knowledge relevant for the Enconverter design. Besides the course already mentioned, it was necessary to deepen the computational formalism of knowledge representation, from the more commonly used in Natural Language to others derived from Knowledge Engineering and Artificial Intelligence, having a narrow relation with Natural Language. In short, the point was that the working group should not have lacks of knowledge neither from the theoretical point of view or that of the computing system development.

4.1.2 Definition of the global computational architecture of the system that would reflect its different modules and define adequately the works to carry out at any time.

4.1.3 Definition of a core of morpho-syntactic rules around which the compiled knowledge of Spanish can increasingly grow in the Enconverter. This core would be based in the use of a free context syntagmatic grammar, and the rules that represent less frequent cases or exceptions would be left for a second phase.

With a view to these premises, we can go on explaining the course of the development of the works covered by this task of the project.

4.2 Development of the works

This section pretends to show the works carried out in the task generically called “Enconverter or Analyzer”. We will try to show not the following-up of a Planning (that as we already explained obeyed to an internal planning) but the conceptual sequence of the works all over the time, that is, seeking for the coherence of the time and contents. The beginning of the project were dedicated to:

4.2.1 Study of the relevant literature and industrial works with a view to our work so as to be able afterwards to draw the lines of work without losing time by following lines of work already run down considering our specific objectives. Some of the books consulted were [BUCH-84], [NILS-87], [WHITE-90], [BOGU-88], [GAZD-89], [LEON-87], [REYLE-88] and [SHIEB-86].

As a result of these preliminary studies, some important conclusions were reached:

- The adequacy of the use of a free context syntagmatic grammar for Spanish, as it is besides a very well known formalism with demonstrated performances in Natural Language.
- The resulting potency and clarity due to the use of feature structures to represent the characteristics associated to each morphologic, syntactic and semantic word.

This was also based on the study of analyzers coming from industrial applications already developed by this working Group [PASO PC 315] [Project sponsored by the European Community and the Spanish Ministry of Industry and Technology] and generated some of our internal documents

4.2.2 Determination of the Enconverter capacities. After many rigorous discussions during which we took into account criteria of rigor, quality and utility, we decided that:

- The system has to cover a wide range of linguistic structures, indeed the fact that it is assisted allows it. (wide range).
- It has to be incremental, that is, it has to permit the addition of new structures if it was estimated to be necessary. (maintainability)
- The system, at least in the initial phase, will be limited to the Spanish language that can be encountered in scientific, journalistic and unspecialized texts.

- We definitively discard the automatic approach for different reasons, among them are the failures of this approach all over the years and also because this would permit to introduce naturally on the market an important help tool for the current professionals of translation.

These considerations have established some very concrete lines in the works carried out that can be summed up as follows:

- A corpus of sentences will not be established in principle.
- The study of the Spanish grammar will have to be global.
- The study of this grammar will be organized in a classical way, that is:

- 1.- Analysis of lexical elements
- 2.- Analysis of phrases.
- 3.- Analysis of complete sentences: simple y complex.

4.2.3 Definition of morpho-syntactic features of the lexical entries. Several grammars of the Spanish language were used as a basis for the study of grammar, among them we can point out [ALAR-94] y [FERN-95]. The result of this work was gathered in the following internal documents:

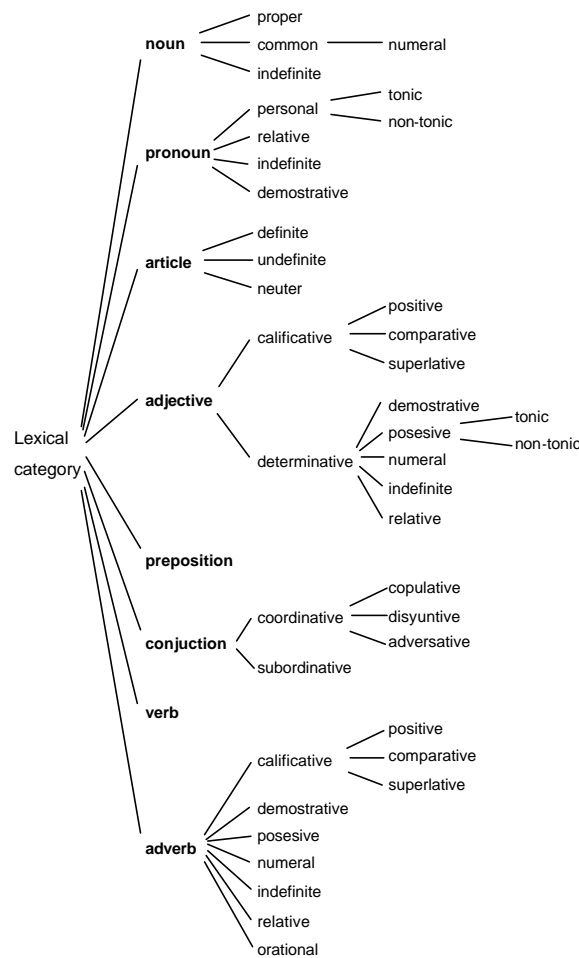


Figure 2: Tree of Lexical Categories and Sub-categories of the Spanish Language

- *Form and Function of the Speech Parts* – [24]
- *The Adverb* – [47]
- *Notes to the document “The Adverb”* – [45]
- *Lexicon y Syntax* – [32]
- *Quantificational Modifiers* – [34]

After making these documents, we proceeded to the definition of the different lexical categories to which a word belongs, as well as the assignation of the grammatical features that determine its behavior. As an example, we can see in Figure 1 the resulting tree of categories:

Each one of the lexical items is characterized, besides by the categories, by a set of grammatical features. These features were represented as a set of three-element tuples Object-Attribute-Value of which we show an example below.

The Common Substantive “perro” would be catalogued as follows:

OBJECT	ATTRIBUTE	VALUE
“perro”	Category	Common substantive
	Morfological pattern	morphological pattern #2
	Gender	masculine
	Number	both
	Animate	yes
	Human	no
	Countable	yes
	Universal Word	dog(icl>animal)

Table 3: Object of the Common Substantive "perro"

The complete list of the grammatical features to be applied to each one of the categories was gathered into the documents:

- *Lexical Categories I* – [35]
- *Lexical Categories II* – [36]

These two documents were an essential entry to define the contents and the form of these ones in the Dictionary, especially for two purposes:

- Detailed specification of the design and content of some of the structures of the Data that constitute the Dictionary.
- Guide for the introduction of the new entries considering the examples that these documents contain.

4.2.4 Representation of the grammatical knowledge. First of all, we tried to find a way of representation of the grammatical knowledge that above all was clear and

easy to understand by all those that would work on the conceptual design of this system, for the lexical as well as syntactic questions.

The lexical knowledge will be represented by directed acyclic graph in which the word is the root node. The branches are labeled with the names given to the grammatical features and the terminal nodes contain the information associated to these features. Here is the graph representing the lexical information associated to the word “perro”:

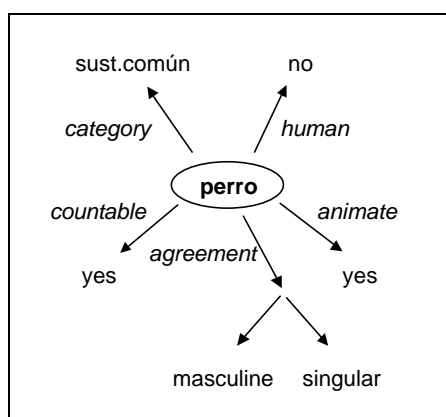


Figure 3 : Lexical Information Graph of the word "perro"

A detailed description of this formalism can be found in the document:

- *Features structures as Directed Acyclic Graphs* – [56]

The categories will be formed by concatenation of ordered sequences of elemental components of a sentence, by these ones we mean the words and locutions. The way of forming a category will obey, besides the previous rule, a set of conditions of application or restrictions depending on the value of the features that characterized these components. In order to facilitate a better comprehension (despite of the future formal representation), rules are expressed as follows:

Example: *a nominal phrase is constituted of a determinant phrase followed by a common noun provided that both components agree.*

The representation of this rule is:

Antecedent:

- The **category** of the **constituent 1** is **determinant phrase**
- And
- The **category** of the **constituent 2** is **common noun**
- And
- **Constituent 1** and **Constituent 2** agree

Consequent:

- The **category** is **nominal phrase**
- And
- The **determinant** is **determinant phrase**
- And
- The **core** is **common noun**

In general, the specification of the value of a feature in a constituent is made indicating the path that leads from the root of the graph to the node that contains this value. Together with the basic operation of matching of a value and another, other operations were added in order to increase the expressive capacity of our rules. This method of representation of the grammatical knowledge is used in the task of compilation of a grammar of the Spanish language.

4.2.5 Compilation of the grammatical rules

In some way, we can say that the content of this section is the one that refers to the final result of the task of this year, or at least the more visible. The philosophy followed because of the impossibility of reflecting the whole grammar of a language in the work of this year and of this task was to take decisions orientated to the efficiency and utility of the work with a view to the future. We can describe shortly the orientations that we have followed to determine a set of rules that reflects a core of the language so as for the grammatical knowledge of the Spanish language. Then we can say that this core should:

- **Be consistent**, that is it should reflect in the rules obtained a set of grammatical assertions free from exceptions. In fact, exceptions have been studied but the corresponding rules have not been compiled, as for that purpose, it will be necessary to make an entire work plan, taking into account the peculiarities of the Spanish language.
- **Rigor**, in the sense of studying systematically all the general cases related to the different syntagms and starting from phrases, that is, parts of the more elemental sentences but that are able to reflect a semantics.
- **Formalisms**. We have defined a very classical formalism that can be extended by the greatest part of the persons involved in Natural Language Systems, without excepting the definition of another formalism in the second part of the project that would be more orientated to clear and efficient computational formalisms as those derived from the object-orientated design.
- **Compactness**, meaning by this one the definition of rules that cover a wide range of cases (according to the values of its variables) and even compact and reflect a higher number of different rules. The compactness permits also an efficient handling of the knowledge base and guarantees the functional validity of the defined core of knowledge. Contrary examples are the exceptions that in general require an individualized processing irrespective of the heuristics used for its handling.

We have studied the following phrases:

- *The Determinant Syntagm* – [41]
- *The Nominal Syntagm and its core* – [27]
- *The Adjective Syntagm* – [33]
- *The Prepositional Syntagm* – [44]
- *The Pronominal Syntagm* – [42]
- *The Verb: Core of the Verbal Syntagm* – [53]

We are going to illustrate the type of rules extracted by two examples. Let's see the rules involved in the recognition of a nominal phrase as "mi perro".

Rule 1: a nominal syntagm combines a determinant syntagm with a common noun. To have agreement, they must coincide in gender and number. The categories involved are:

- (Nominal_phrase): *Mi perro*
- (Determinant_Phrase): *mi*
- (Common_Noun): *perro*

The rule to be applied is:

Antecedent:

- The **category** of the **constituent 1** is **determinant phrase**
- And
- The **category** of the **constituent 2** is **common noun**
- And
- **Constituent 1** and **Constituent 2** agree

Consequent:

- The **category** is **nominal phrase**
- And
- The **determinant** is **determinant phrase**
- And
- The **core** is **common noun**

Rule 2: A determinant phrase is formed by a possessive determinative adjective.

Antecedent:

- The **category** of the **constituent 1** is **possessive determinative adjective**

Consequent:

- The **category** of the resulting phrase is **determinant phrase**
- And
- Its core is the **possessive determinative adjective**

We can illustrate the application of both rules showing the graph of the resulting phrase constituted by the information provided by the Dictionary. For the application of the second rule we assume that a word like "me" has been catalogued in the Dictionary as follows:

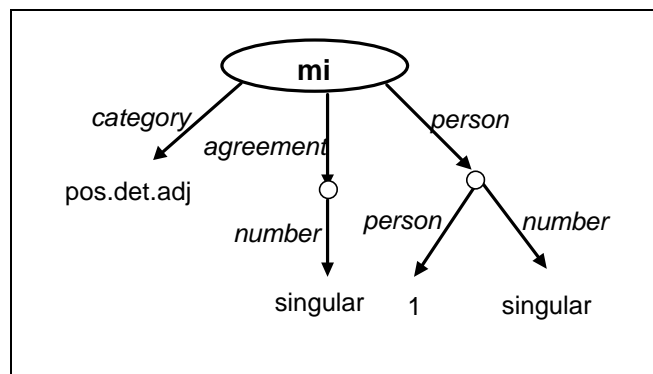


Figure 4: Graph representing the features of the word "mi"

And therefore the application of the second rule will produce a new determinant phrase with the following structure:

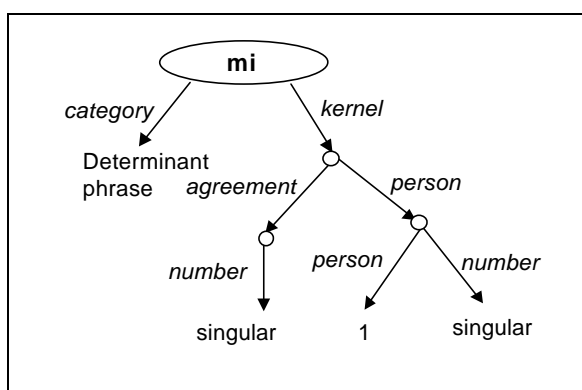


Figure 5: Graph representing the features of the determinant phrase

Finally, the application of rule 1 on the determinant phrase just form and on the noun “perro” –which graph has been previously shown- produces the phrase “mi perro”:

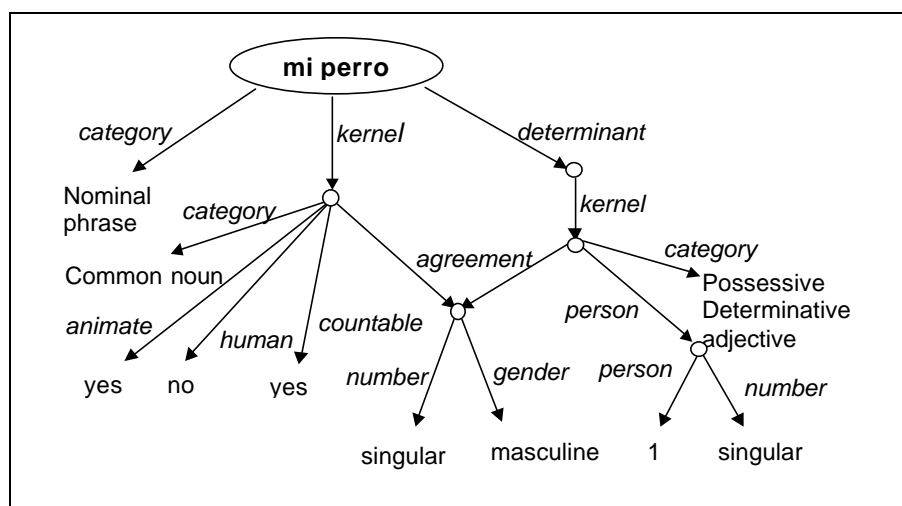


Figure 6: Graph representing the features of a nominal phrase

The totality of the rules produced are gathered in an internal document “*Rules for the definition of sub-sentential phrases*” [52] and is enclosed as Annex 3.

4.2.6 Global architecture of the enconverter. The architecture or if you want the global conception of the Enconverter has to obey to structures able to assume the global specifications of the system.

They can be summed up as follows:

- 1.- The simple or complex sentence is the basis of the analysis.
- 2.- A morpho-syntactic analysis of the sentence has to be made.
- 3.- A semantic analysis of the sentence has to be made and it has to be represented.

4.- This set of results should be displayed to the user of the Enconverter and allow him to make changes on the results proposed by the system. That is, the user validates the results of the Enconverter on-line.

5.- Once validated by the user, it should traduce the proposal automatically to UNL.

In principle the Enconverter user should be a linguist or translator. In that sense a rapid and very professional result is assured and the absence of errors in the transcription of a text to UNL too.

We consider the possibility that the user can add items to the dictionary from this interface but this will be contemplated in the second year specifications.

We can explain it better by a graphical scheme:

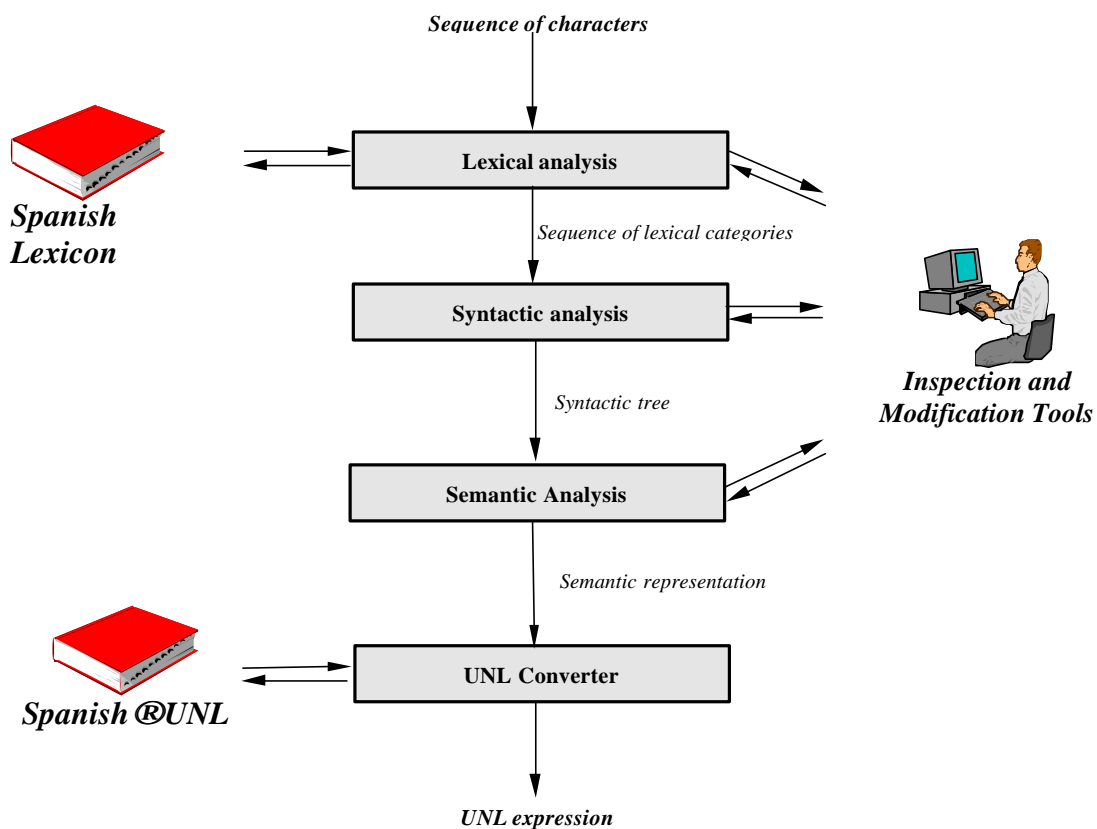


Figure 7: Global Architecture of the Analyzer

The simple idea of the functioning is that the user can introduce a text in Spanish whether typing it directly or using a file. The system starts to work on this text and shows in a sequential way a series of results to the user, who, in turn and dynamically, can change as a professional what he/she thinks should be changed. Once the user is totally satisfied with what the system reflects, he/she gives to the system the instruction of translating to UNL. It is then when the Enconverter will interact with the UWs. The system will be complemented with a series of tools of help and handling of the documents.

CONCLUSIONS

The thorough study of this module called Enconverter and dedicated to the language analysis –one of the greatest sources of failure in the implantation of the Linguistic Engineering at the industrial level-, has increased our consideration as for the role it plays in this project. Indeed, the best generator will give mediocre results if UNL expressions associated to a text are not strictly correct. This module is then essential for the UNL project to have an industrial exploitation or to be useful for the society in a near future.

This module has a part which have to be strictly realized by each country and another part that can be done in a common way, or at least making good use of UNL Specifications. The individual part is obviously that which corresponds to the grammatical analysis of each language. The part that could be common but not in a strictly necessary way, is that which provides the internal processes of the system to access the Universal Words (UW) and the UNL expressions. A third part that could be agreed on or be independent for its designer, is that related to the user interface and its performances.

The study of these questions consumed part of the time of work during this year. This has resulted into an architecture that will have to be determined in a detailed way in the second year. This second year should then determine a set of formal specifications of the Enconverter as a previous step to the computational development immediately after.

Another part, -the most important-, was dedicated to the study of the grammatical characteristics of the Spanish language to generate a core of rules that could be quantitatively increased but that should regulate some minimum capacities which should serve as a validating component of the Enconverter already developed. Without going in more details, we can affirm that this module is the more critical of the three essential modules of the UNL project. As for the Dictionary, we can say that, once validated, whatever its size, the only thing to do will be to increase its contents and capacity. The generators can be debugged and improved but they cannot show their quality if the UNL expressions –generated by the Enconverter–, are not good. The analysis of the Language has been the hobby-horse of the researchers of the Natural Language area for ages. The best results were obtained when the context is limited and defined. But being free, the analysis and the comprehension of a text are the more *intelligent* part of those systems. Hence the entry of the Artificial Intelligence in this world. New methods, new ideas and our own experience of more than ten years in industrial products make us somewhat optimistic, although everything must be proved and we will dedicate our efforts to it.

5. DICTIONARY

Any system based in the idea of Interlingua disposes of a dictionary used as lexical source between the language analyzers and generators. The engagement for the first year of work was the building of a Database that should be composed by 100.000 pairs with the form :

(Spanish inflected form, Universal Word)

Naturally, this dictionary should be equipped with a set of tool of access, maintenance and integration with the Deconverter format proposed by the IAS. Having in mind this basic idea, we can go on describing the tasks carried out.

5.1 Development of the Dictionary

5.1.1 PRELIMINAR DESIGN. By this design, we mean all the tasks that reflect the recollection of needs and the conception of the system at the architectural as well as computational levels. Unlike the analyzer which did not require a computational implementation in the first year, the dictionary has to be implemented on pain of not visualizing the work. This point of view had an influence on two aspects that would define jointly how the dictionary should be.

- **Contents**, that is the words that should compose it and proportion of these words depending on their more or less frequent use in Spanish.
- **Sub-systems to which it should serve**, that is, the Dictionary should support the activities of the Enconverter as well as of the Deconverter.

We are going to comment the line of work followed by these two questions.

5.1.2 CONTENTS: After numerous consults and research, we designed a thematic tree that will be the guide for the contents of the Dictionary in the short and medium term. This thematic tree can be see in Figure 8. Another criterion of contents was the incorporation of all the words derived from the corpuses indicated by the IAS, specifically and currently all the words derived from the “Tower of Babel” and of all the sentences of the corpuses delivered by the IAS are obviously introduced on pain of not being able to demonstrate the viability of the Deconverter. Besides the thematic tree, we made a study of proportions of the words that should be present to assure that we were building a balanced dictionary that would guarantee the basic idea of design of a free-context global system.

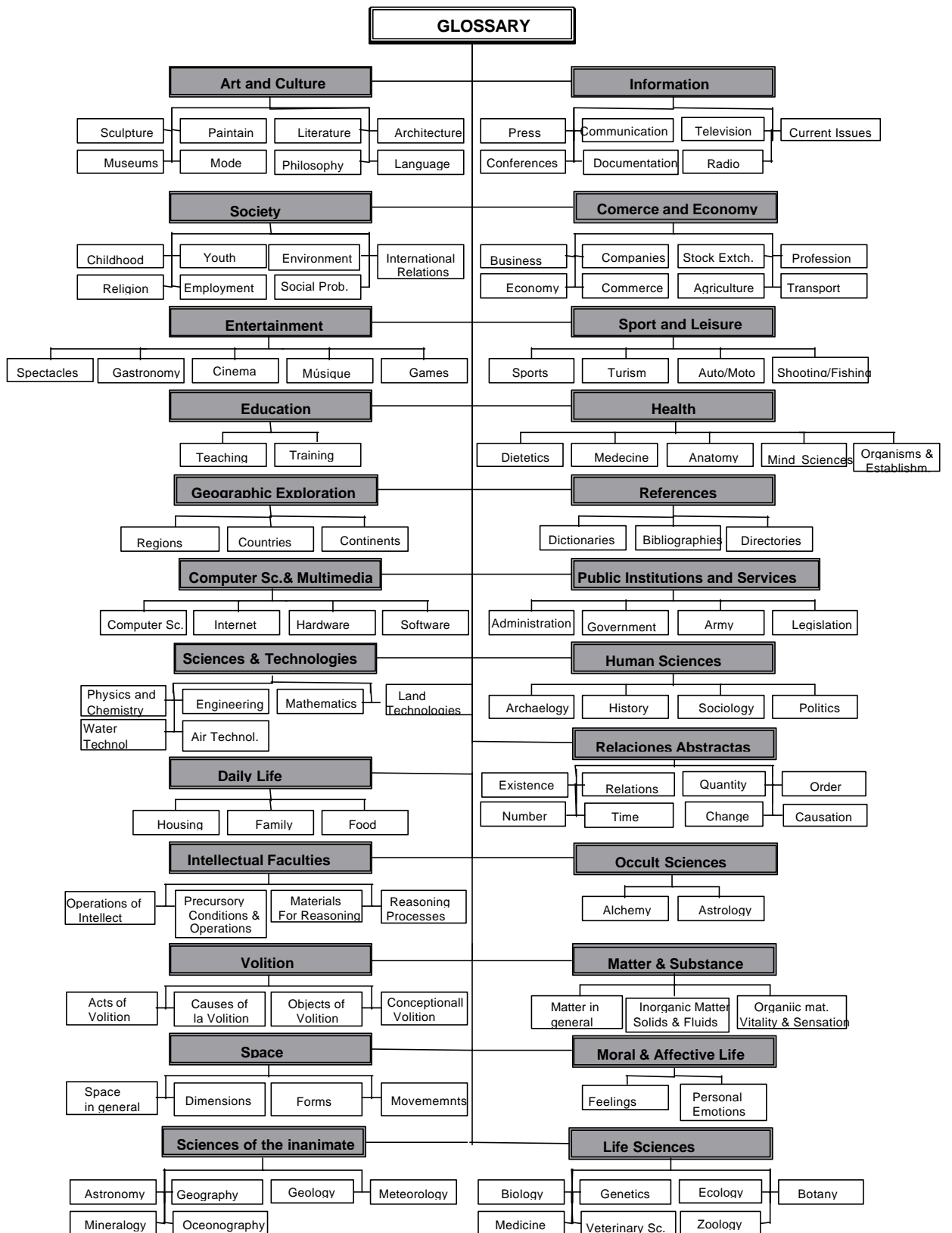


Figure 8 : Thematic tree of the Lexicon

You can see in Table 4 a table of the content proportions which will constitute a reference as the contents of the dictionary increases. A great part of these studies were based on the study of the thesauruses and particularly Roget's

Thesaurus of which an exhaustive document of study was made [ROGET-97]. We will comment the content of the Dictionary at the formal level after this section explaining the type of information associated to this Dictionary content.

Lexical Field	Proportion
Art and Culture	10 %
Life Sciences	10 %
Sciences of the Inanimate	4 %
Human Sciences	3 %
Occult Sciences	1 %
Sciences and Technologies	15 %
Commerce and Economy	8 %
Sport and Leisure	3 %
Education	3 %
Entertainment	4 %
Space	2 %
Geographic Exploration	2 %
Intellectual Faculties	2 %
Information	4 %
Computer Science and Multimedia	1 %
Public Institutions and Services	2 %
Matter and Substance	2 %
References	1 %
Abstract Relations	4 %
Health	10 %
Society	4 %
Daily Life	2 %
Moral and Affective Life	2 %
Volition	1 %
TOTAL	100 %

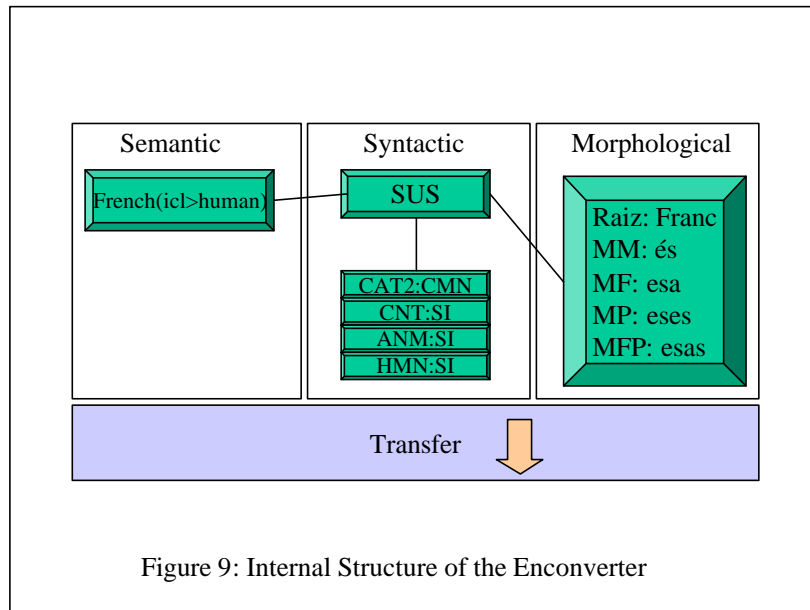
Table 4: Dictionary Content Proportions

5.1.3 INTERNAL STRUCTURE : Each Universal Word (UW) by its essentially semantic content belongs to a Semantic Block. To each Universal Word is associated a series of contents that form what we call the Syntactic Block. The content of the Syntactic Block is based on the Syntactic information related to how this word is used in Spanish. This syntactic information is composed by :

- Grammatical **Category** of each word.
- A set of linguistic **features** among them are included the verbal pattern, that define any type of characteristic with syntactic or semantic information associated to each word.

Finally, to each Syntactic Block (there can be more than one per Universal Word) corresponds at least one Morphologic Block in which the corresponding Spanish word is introduced. In each Morphologic Block, we introduced the information contained in the so-called Conjugating Tables, in the case of verbs it will be each one

of the verb endings, and in the case of nouns and adjectives, it will be the different combinations of Gender and Number. An illustration of this can be see below in Figure 9



For operative reasons and for an initial control (modularization of the work), we provisionally designed an additional Database to store the Spanish verb endings. The content of this table was built in basis of the Classifications of Conjugating Models [MATEO-84] up to a total of 90. That is, there are 90 different ways of conjugating a verb in Spanish. A number or index was assigned to each of the 56 verbal forms of a Spanish verb, these numbers will be used to identify these forms in any case. A total of 174 root tables corresponding to the 90 Models were built, indeed various roots can correspond to a determined model. You can see below a series of tables that display the list of the 56 Spanish verbal forms.

Mode	Tense	Person	Nº	
Infinitive	Infinitive		1	
	Gerund		2	
	Participle		3	
Indicative	Present	1ª pers.sing.	4	
		2ª pers.sing.	5	
		3ª pers.sing.	6	
		1ª pers.pl.	7	
		2ª pers.pl.	8	
		3ª pers.pl.	9	
		Imperfect	1ª pers.sing.	10
			2ª pers.sing.	11
			3ª pers.sing.	12
1ª pers.pl.	13			
2ª pers.pl.	14			
3ª pers.pl.	15			
Preterite	1ª pers.sing.	16		
	2ª pers.sing.	17		
	3ª pers.sing.	18		
	1ª pers.pl.	19		
	2ª pers.pl.	20		
	3ª pers.pl.	21		
Future	1ª pers.sing.	22		
	2ª pers.sing.	23		
	3ª pers.sing.	24		
	1ª pers.pl.	25		
	2ª pers.pl.	26		
	3ª pers.pl.	27		
Conditional	Simp. Cond.	1ª pers.sing.	28	
		2ª pers.sing.	29	
		3ª pers.sing.	30	
		1ª pers.pl.	31	
		2ª pers.pl.	32	
		3ª pers.pl.	33	

Mode	Tense	Person	Nº	
Subjuntive	Present	1ª pers.sing.	34	
		2ª pers.sing.	35	
		3ª pers.sing.	36	
		1ª pers.pl.	37	
		2ª pers.pl.	38	
		3ª pers.pl.	39	
		Preterite	1ª pers.sing.	40
			2ª pers.sing.	41
			3ª pers.sing.	42
1ª pers.pl.	43			
2ª pers.pl.	44			
Future	1ª pers.sing.	46		
	2ª pers.sing.	47		
	3ª pers.sing.	48		
	1ª pers.pl.	49		
	2ª pers.pl.	50		
Imperative	Present	3ª pers.sing.	51	
		2ª pers.sing.	52	
		3ª pers.sing.	53	
		1ª pers.pl.	54	
		2ª pers.pl.	55	
		3ª pers.pl.	56	

Table 5 : The 56 Spanish Verbal Forms

We can also see in Table 6 an example of the tables derived from the models mentioned above.

We will not include the tables or lists of features handled by the system (except if you require them). The reason is just that we do not want to surcharge this document with tables of internal use at the present time. Anyway, this table can be found in the Document “*Features of the UNL Dictionary*” [30].

The content of the Dictionary itself is enclosed in Annex 4. In this Annex, we explain how the information saved on the enclosed CD-Rom is. This CD-Rom is structured as follows :

- A) Verbal endings.
- B) Words (implicit 101.902 pairs)

[iendo]{} desinencia " TB7" (MINF,GER);	[iréis]{} desinencia " TB7" (IND,FUT,2PL);
[ido]{} desinencia " TB7"	[irán]{} desinencia " TB7" (IND,FUT,3PL);
(MINF,PAR,MAS,SIN);	[iría]{} desinencia " TB7" (CON,CDSM,1SG);
[ida]{} desinencia " TB7"	[irías]{} desinencia " TB7" (CON,CDSM,2SG);
(MINF,PAR,FEM,SIN);	[iría]{} desinencia " TB7" (CON,CDSM,3SG);
[idos]{} desinencia " TB7"	[iríamos]{} desinencia " TB7"
(MINF,PAR,MAS,PLU);	(CON,CDSM,1PL);
[idas]{} desinencia " TB7"	[iríais]{} desinencia " TB7"
(MINF,PAR,FEM,PLU);	(CON,CDSM,2PL);
[ir]{} desinencia " TB7" (MINF,INF);	[irían]{} desinencia " TB7" (CON,CDSM,3PL);
[o]{} desinencia " TB7" (IND,PRES,1SG);	[a]{} desinencia " TB7" (SUB,PRS,1SG);
[es]{} desinencia " TB7" (IND,PRES,2SG);	[as]{} desinencia " TB7" (SUB,PRS,2SG);
[e]{} desinencia " TB7" (IND,PRES,3SG);	[a]{} desinencia " TB7" (SUB,PRS,3SG);
[imos]{} desinencia " TB7" (IND,PRES,1PL);	[amos]{} desinencia " TB7" (SUB,PRS,1PL);
[ís]{} desinencia " TB7" (IND,PRES,2PL);	[áis]{} desinencia " TB7" (SUB,PRS,2PL);
[en]{} desinencia " TB7" (IND,PRES,3PL);	[an]{} desinencia " TB7" (SUB,PRS,3PL);
[ía]{} desinencia " TB7" (IND,PTIM,1SG);	[iera]{} desinencia " TB7" (SUB,PTI,1SG);
[ías]{} desinencia " TB7" (IND,PTIM,2SG);	[ieras]{} desinencia " TB7" (SUB,PTI,2SG);
[ía]{} desinencia " TB7" (IND,PTIM,3SG);	[iera]{} desinencia " TB7" (SUB,PTI,3SG);
[íamos]{} desinencia " TB7" (IND,PTIM,1PL);	[iéramos]{} desinencia " TB7" (SUB,PTI,1PL);
[iais]{} desinencia " TB7" (IND,PTIM,2PL);	[ierais]{} desinencia " TB7" (SUB,PTI,2PL);
[ían]{} desinencia " TB7" (IND,PTIM,3PL);	[ieran]{} desinencia " TB7" (SUB,PTI,3PL);
[í]{} desinencia " TB7" (IND,PTID,1SG);	[iere]{} desinencia " TB7" (SUB,FUTR,1SG);
[iste]{} desinencia " TB7" (IND,PTID,2SG);	[ieres]{} desinencia " TB7" (SUB,FUTR,2SG);
[ió]{} desinencia " TB7" (IND,PTID,3SG);	[iere]{} desinencia " TB7" (SUB,FUTR,3SG);
[imos]{} desinencia " TB7" (IND,PTID,1PL);	[iéremos]{} desinencia " TB7"
[isteis]{} desinencia " TB7" (IND,PTID,2PL);	(SUB,FUTR,1PL);
[ieron]{} desinencia " TB7" (IND,PTID,3PL);	[iereis]{} desinencia " TB7" (SUB,FUTR,2PL);
[iré]{} desinencia " TB7" (IND,FUT,1SG);	[ieren]{} desinencia " TB7" (SUB,FUTR,3PL);
[irás]{} desinencia " TB7" (IND,FUT,2SG);	[e]{} desinencia " TB7" (IMP,2SG);
[irá]{} desinencia " TB7" (IND,FUT,3SG);	[id]{} desinencia " TB7" (IMP,2PL);
[iremos]{} desinencia " TB7" (IND,FUT,1PL);	

Table 6: Example of table derived form the verbal models

5.1.4 TOOLS: The introduction of the data inherent to the Dictionary as well as the introduction of the words themselves, or even the maintenance and modifications required previously the development of a series of tools that we will describe below :

- 1) **UNLDic:** This tool is the main one as it is that which permits to introduce, delete or change the entries of the Dictionary. The program is written in C Language and runs as a client of the Database Server. An especially useful element is that permitting to introduce verbal ending tables. It also permits to consult or modify the tables already introduced, It can run on various types of UNIX workstations and allows various operators to update the UW Database or the Spanish words. Its user interface is a text interface and can be used through a connection to the server by a "telnet" application. The aspect and characteristics of the interface can be seen below in Figure 10.

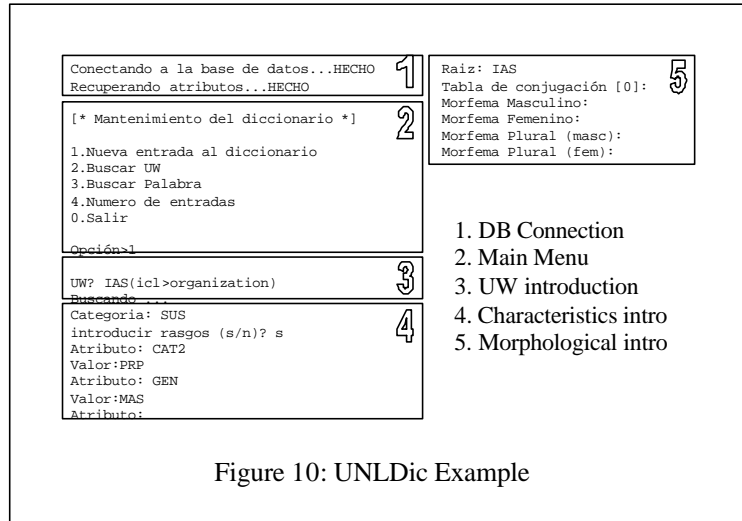


Figure 10: UNLDic Example

In this example, we can see the execution of the UNLDic application:

1. Connection with remote database (hosted in dictionary server)
2. Main menu. The option 1 permits to enter new entries and the option 2 permits to modify these entries. (option 1 selected)
3. Introduction of UW (IAS(icl>organization)) and searching in the database for the existence of this UW.
4. Linguistic features introduction, first we introduce the name of the feature and then the value.
 Category: SUS
 Subcategory: PRP
 Gender: MAS
5. Morphologic introduction (IAS has no gender-number variation, so the only field to fill is Raíz - root-)

- 2) **UNL SINOM**: The objective of this tool is to permit the introduction and storing of conceptual synonyms in the Dictionary, by the simple association of the information already existing on a UW with that which is introduced. By conceptual synonyms we mean the UW that have the following characteristics :

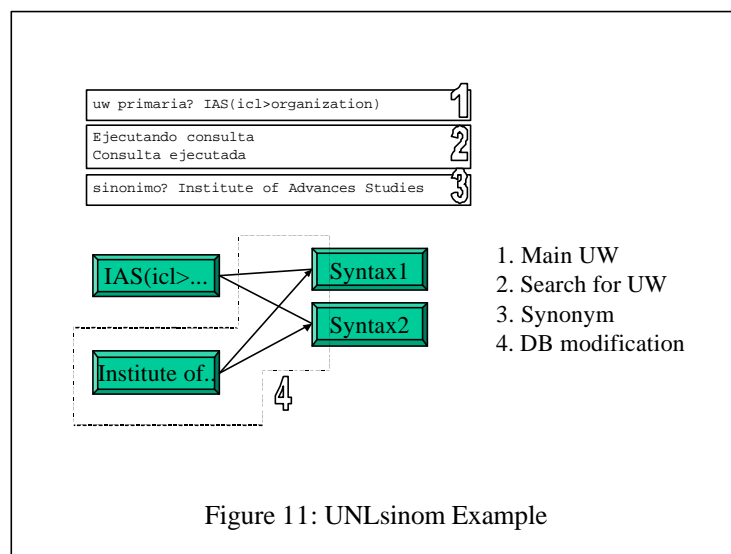


Figure 11: UNLsinom Example

- They have a unique meaning in Spanish, as it is the case for UWs based on the same Headword with different modifiers. For example: **book(fld>hotel)** and **book(equ>reservation)**.
- They have the same representation in Spanish.
- They have the same features in Spanish.

The utility of this tool is that it permits to introduce quickly the UWs that appear under different forms in the corpuses. We can see an example below in Figure 11. In this example we can see how the tool creates a new semantic block and relates it to the syntactic blocks.

3) **DBCohere** : This tool was built to allow the developer to assure the validity of the contents of the Database. One of the objectives is to eliminate from the Database the references to not-existing or not-referred Blocks. It also permits to delete tables whose data are not complete or have not reciprocal sense. Basically, they are two programs written in C and called **Modif** for the correction of indexes, and **Limpia** for the elimination of tables. These applications are executed periodically when the Database is not being used.

4) **WEBACCESS** : This tool permits to access to the Database through the WWW. It is a program written in C and called **busca.cgi**.

It permits to search in the Database for UWs or for Spanish words and it shows the information associated. The program accesses to the verbs of the Database to conjugate them and add the enclitics so as to check if a word is stored in the Database. The aspect of the interface and example is shown below in Figures 12 and 13.

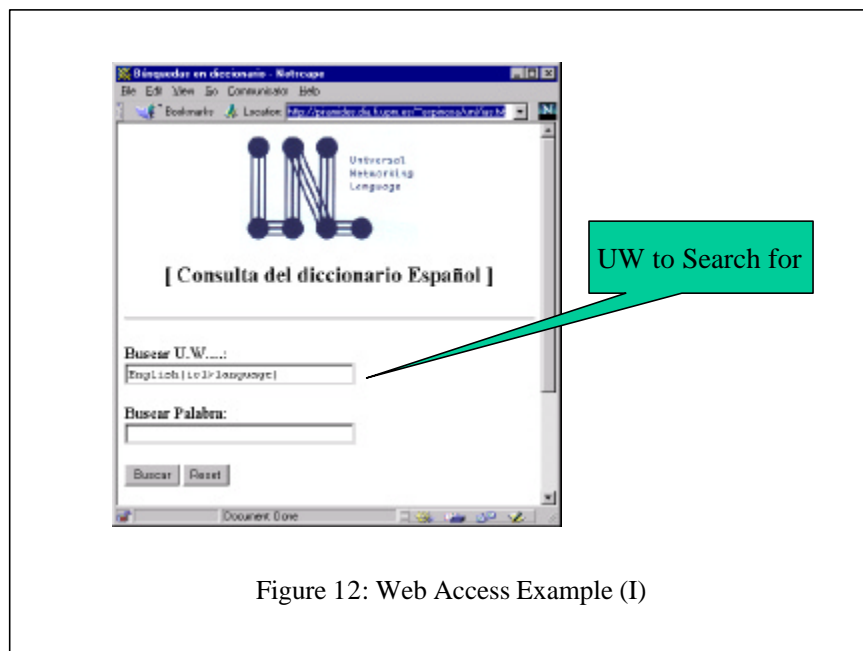
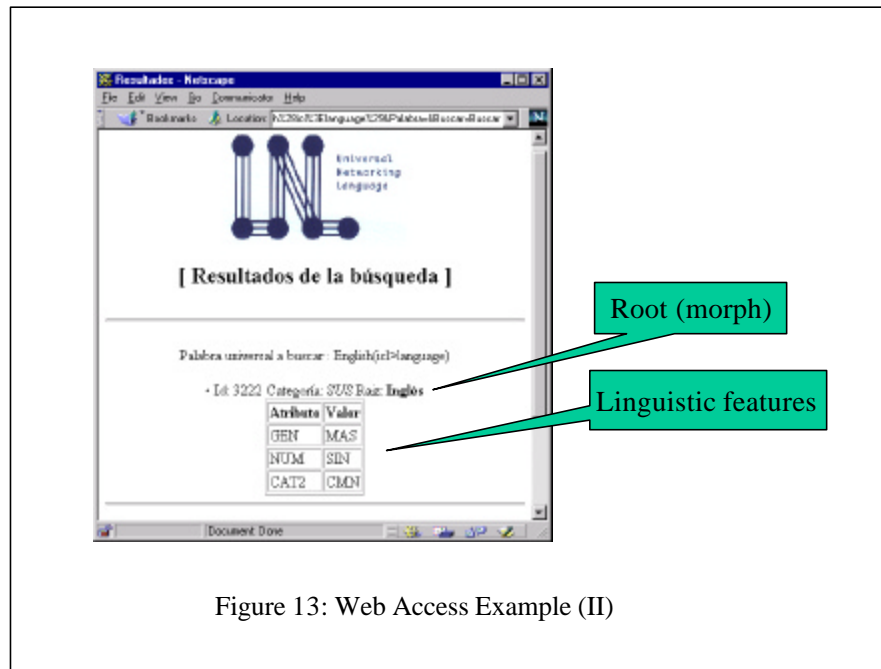


Figure 12: Web Access Example (I)



- 5) **WORD FILTERING** : This tool aims to avoid repetitions in the selection of entries of the Dictionary and it checks that the words of a list were correctly introduced. This tool permits to introduce words from public lists with a great efficiency as it permits to know which are the ones that are already introduced and then to do the corresponding selection.
- 6) **SUBSTAN** : This tool written in Prolog permits the semi-automatic introduction of nouns in the dictionary. The application has two parts. One receives a list of nouns and it deduces automatically the information associated to them and shows it to the user so that he checks if it is correct. Another application written in C introduces these words in the Database once they have been checked by the user.
- 7) **TRANS** : The objective of this application is to transform the content of the Dictionary to the Deconverter format so that it can be used by this one. The DECOL format consists in a file in which each line represents a Spanish word (or a set of them in the case of verbs). The semantic, syntactic and morphologic information contained in the Dictionary is caught by the transference program to build the entry to the Deconverter in this format. This application generates in turn additional information to complete the format, for example it adds information on the verbal tense, number, person or mode to the irregular verbs. We explain it in more details, step by step, in the following :
 - First, all the morphological information about verbal endings is dumped. This information covers all the possible endings for regular and irregular verbs in Spanish.
 - Next, all the morphological blocks are indexed to serve as a guideline to the dictionary entries. We use the morphological blocks as a primary token due to the headword required in the DeCol format entry.
 - Using the information contained in the morphological block, the application compounds the headwords derived from it. That is, it builds all the possible variations in gender and number.

- The next step is to collect the syntactical blocks associated to the selected morphological block. Then it dumps their features linked to each syntactical block, mapping their names to the final format.
- The last step consists in retrieving the UW corresponding to the morphological and syntactical block.

D) WRITING OF THE DICTIONARY : This activity is obvious in this task but it had to be carried out to comply with the specifications of the contract and to be able to generate as well as analyze. Considering this aspect, we greatly reached the objective fixed as the figures obtained are :

Number of Pairs: 101.952

Number of Semantic Blocks: 9.807

Number of Morphologic Blocks: 11.328

Then the quantitative objective has been fulfilled.

As we said before, Annex 4 and the information of the CD-ROM enclosed contain this material.

CONCLUSIONS

- The dictionary supposed a considerable work and received considerable inputs for its design from the Analyzer as well as Deconverter teams.
- The architectural design cannot be considered as definitive although it is currently enough operative thanks to the transference programs, particularly the DECOL.
- The increase in the number of words in a second phase is very important if we want to reach the objective of being able to deal with any kind of language whatever the context.
- We have generated a large number of internal technical documents to assure the following-up of the design and implementation.

6. DECONVERTER

6.1 Study of the Deco (software delivered by the IAS/UNU)

This subtask started to be carried out after the delivery of the Deco software and the associated documentation (example of execution, etc.) by the UNL Center, during the regional meeting held in Pisa (Italy). Although there was some delay in the starting date of this task, the resources employed were reduced because of the simplicity of the software delivery and its easy understanding.

The results of this study can be summed up in the following points:

- Simplicity in the use of the inference mechanisms for the transformation of a network (UNL representation) into a list (final representation in the target language).
- Environment for the interaction with the inference, execution and debugging engine, inadequate for the development tasks.

Document produced: “*Study of the Deconverter*” – [23]

6.2 Design of the rules:

First of all, in March, we made a study and analysis of the first corpus delivered by the UNL Center, called “*Tower of Babel*” (with approximately 100 expressions) [UNL-96]. These studies not only researched the understanding of the Language, but also to observe the adequate way to code expressions in Spanish so as to write a proper test corpus (based on simple expressions).

At the same time, we started to write a series of grammatical reports that would approach the final representation of the Spanish language to UNL conceptual representation. The main objective was to point out the necessary information that would have to be part of the Dictionary contents, as well as the identification of grammatical structures. Those reports comprised the following aspects of the Spanish grammar (see Bibliographic references on Spanish Grammar):

- **Verbal Syntagm.**
- **Nominal Syntagm.**
- **Subordination.**
- **Coordination.**
- **Elements of relation.**
- **Prepositional Syntagm .**

Regarding the Design of the Generator, two fundamental phases were taken into account:

- 1) Functional design, independent of the tool or programming language that would implement it physically.
- 2) Detailed design that would determine the implementation directives and the final form that would have the Functional Design modules.

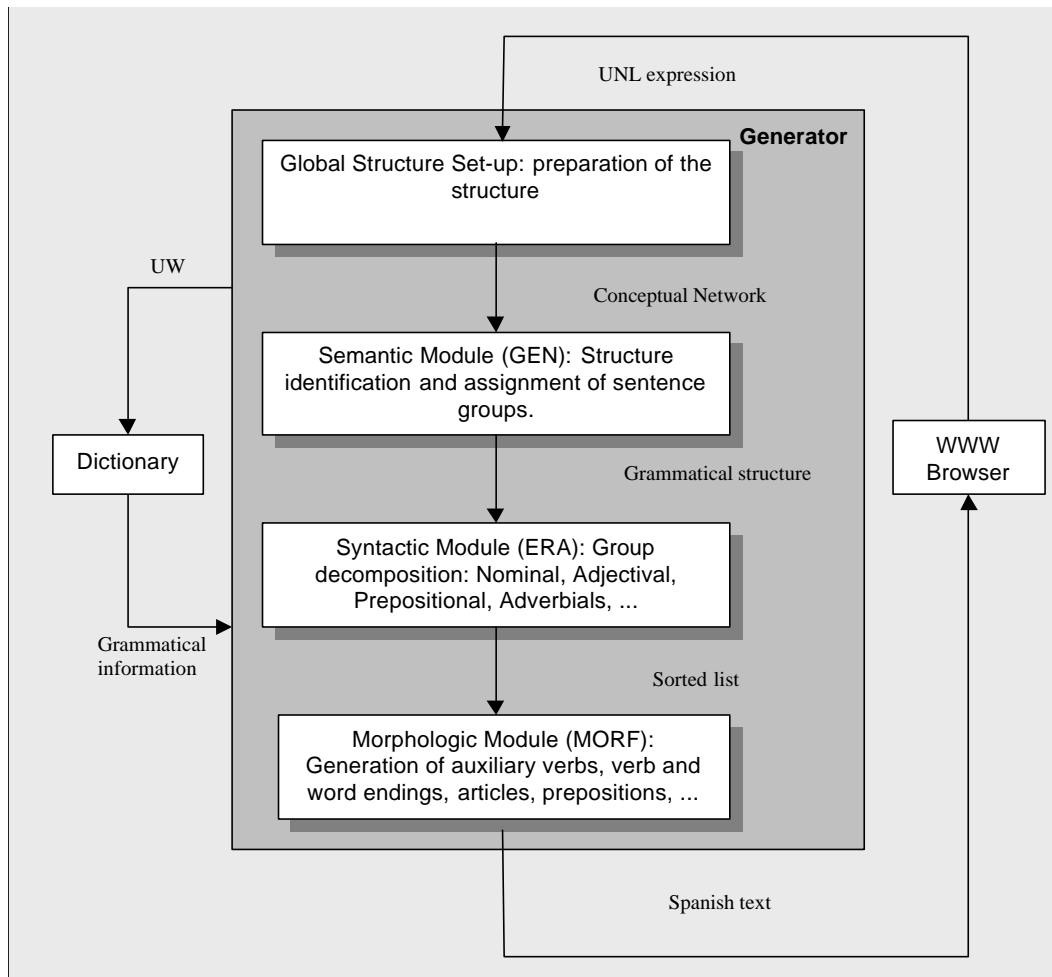


Figure 14: Functional Design of the Generator

The scheme of the functional design can be observed in Figure 14. The interaction with the WWW Browser was obviated as it was not part of the first year objectives of this project, whereas the interaction with the consults to the dictionary is conceptually an access to lexical data base through an Universal Word.

The first step of preparation of the structure would be in function of the input provided, depending on the format implanted: HTML extended by the UNL Center, or directly the relations in plain text. We considered that this first module was part of the interface with the Browser so we did not carry out any work on it.

Before starting the Detailed Design, we made a study to evaluate the different existing alternatives for the utilization of the tools delivered by the UNL Center:

- Complete rejection (that is, independent implementation),
- Partial utilization (several passes on the DeCo itself, or implementation of some modules in an independent mode), and
- Integral use of the tools.

This evaluation brought us to chose the third option, that is the integral use of the tools provided by the UNL Center (DeConverter and Dictionary Builder) [DECO-97a] [DECO-97b], as well as the adaptation to their work formats. This decision was

made for practical reasons (simplicity, easy learning, feasibility for the resolution of simple grammars, etc.) as well as for political reasons: to maintain a homogeneous system between the different working groups (solving of problems together, uniform final product). We Will explain the different modules below:

6.3 Semantic Module

In Figure 15, we show the internal structure of the semantic module. Basically, it identifies the global structure of the sentence, at the phrase level (composition, passive forming, impersonality, etc.), as well as the descriptive level of the sentence: subject (if there is one), compulsory complements, no compulsory complements. That is the identification is made at the highest level possible (leaving aside the context of the sentence, etc.)

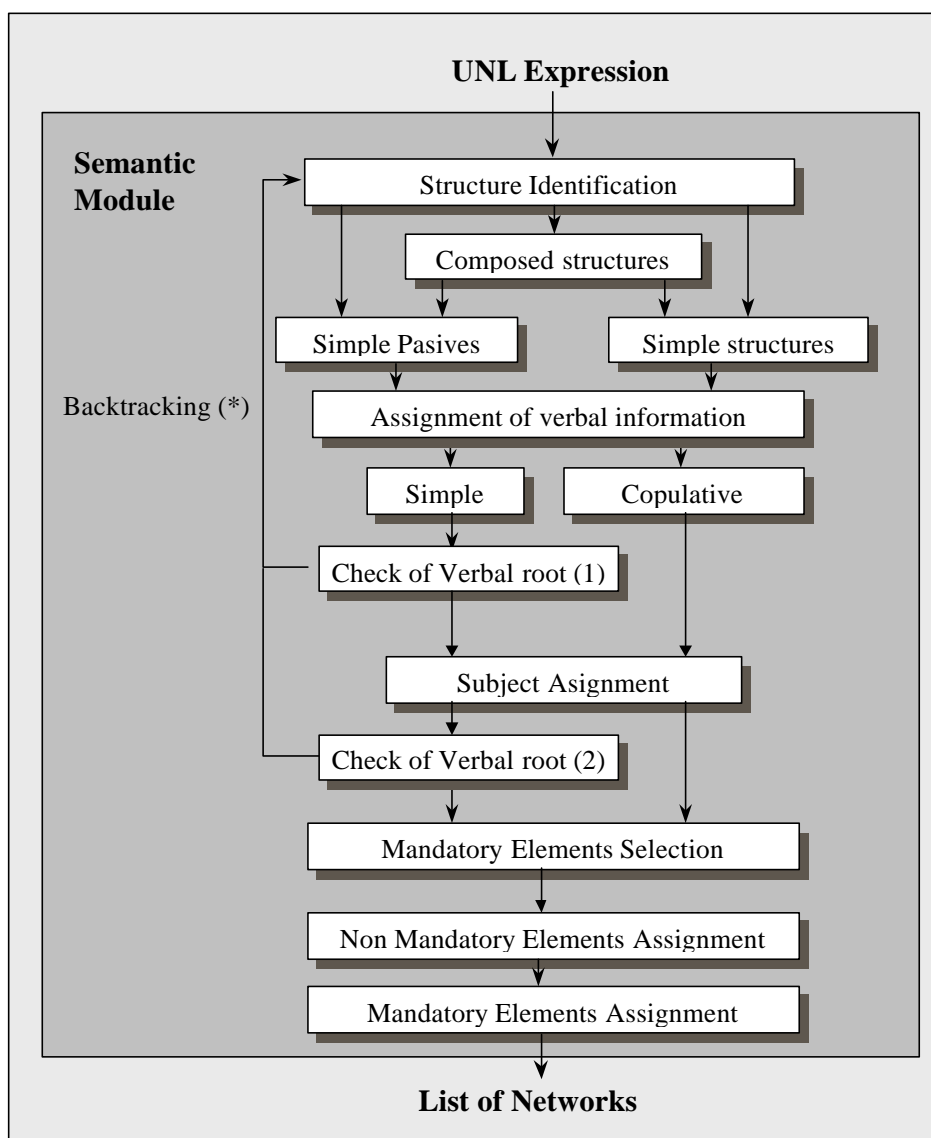


Figure 15: Detailed Semantic Module

The backtracking is used in this module to select the correct root, in the case of irregular verbs, for the verbal aspect, tense, mode, intention (1) as well for the number and person (2). The use of the backtracking to these ends is conditioned by the way of codifying the irregular verbs in the DeCo format dictionary: each root contains a table that identifies the different irregularities of the endings according to the mode, tense, person and number, so that a same Universal Word (UW) can have various different roots. On the other end, another alternative of design was raised, it was the inclusion of a unique headword for each UW, creating tables that would contain not only the information of the irregularities of the endings, but also those of the root.

This last alternative for the design of the DeCo Format Dictionary has not been totally rejected, as, in principle, it would eliminate the backtracking mechanisms for the correct root selection, but it would produce an extra load of functionalities on the Dictionary that would have to correspond to the DeCo design.

The output this module gives to the following is a list of sub-networks, in which the roots (input nodes) of these sub-networks are the elements that they have identified in the node execution. An example of the output given by this module is shown in Figure 16.

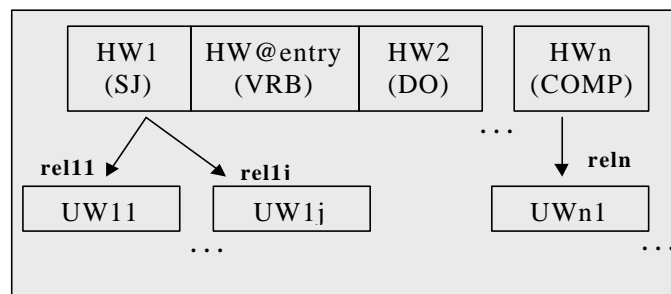


Figure 16: Example of a sub-networks list, output of the Semantic Module

6.4 Syntactic Module

In Figure 17, we show the composition and the groups of rules that are used to process the different syntagms. The execution of one or other group of rules is determined by the type of root node of the network selected. We must point out that the list is went through from the left to the right (from the start to the end), and that a special control is maintained on the groups of relations defined as SCOPE.

Regarding the relative, conditional clauses, etc., in short, any verb (excepting the principal verb of the sentence) that belongs to the structure identified in the previous module, will have the processing that is appropriate in each case: generation of not personal forms (participle, gerund and infinitive) or generation of a relative pronoun. But it will always jump on a concrete point of the semantic module to process the sentence according to its lacks: if it has no subject, if the verbal tense and mode are not indicated, etc.

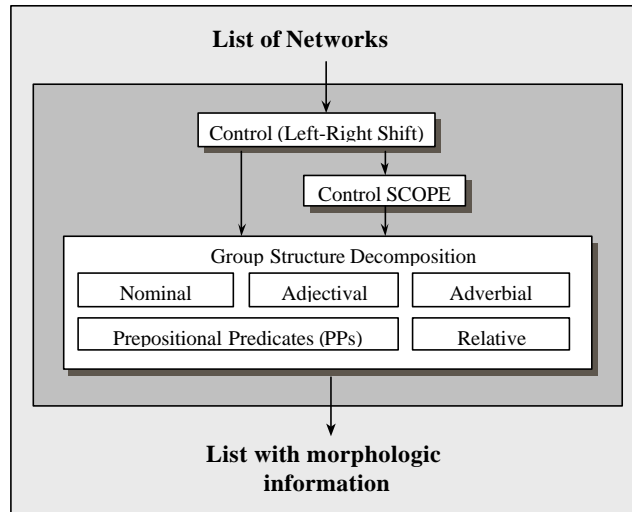


Figure 17: Detailed Syntactic Module

The output of this module produced is a “plain” ordered list to which the only thing lacking are morphologic features, that is, there is not any node hanging form the list, all them should have been processed.

6.5 Morphologic Module

Although this module is conceptually separated from the previous one, the way to carry out the control follows the same scheme as in the Syntactic Module. So that, considering the politics of modification of the DeCo list, following a “Update and Back” scheme (that is modification and shift to the left), it permits to implement the control of both modules with the same rules of shift and processing of SCOPEs.

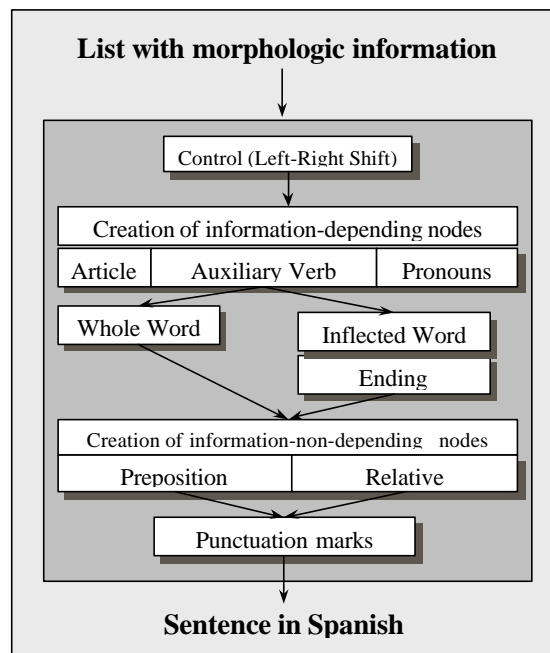


Figure 18: Detailed Morphologic Module

The morphologic generation has two aspects, the generation of nodes that depend on the information contained in a basis node and that of nodes that do not depend on any information. Among the first one, for example, there are the auxiliary verbs, articles, pronouns and of course the endings of each word. In the second group, in which the proper information of the node is not considered, we can find the prepositions, relatives and punctuation marks of any kind.

6.6 Implementation

The implementation of the groups of rules described before, together with the integration of the different modules have their starting point in the definition of an interface between the General Dictionary and the Generator itself. This interface responds to the specific needs for the Generation to be defined and identified from the general Dictionary (Lexicon). This interface also defines how is structured the information of the Dictionary (Dumps to the DeCo Format) and how it will be used inside each module (determining to the highest level if it will be used for the generation of articles, the global structure, etc.). The integration of the rules with the DeCo and the rest of the architecture is in Figure 19.

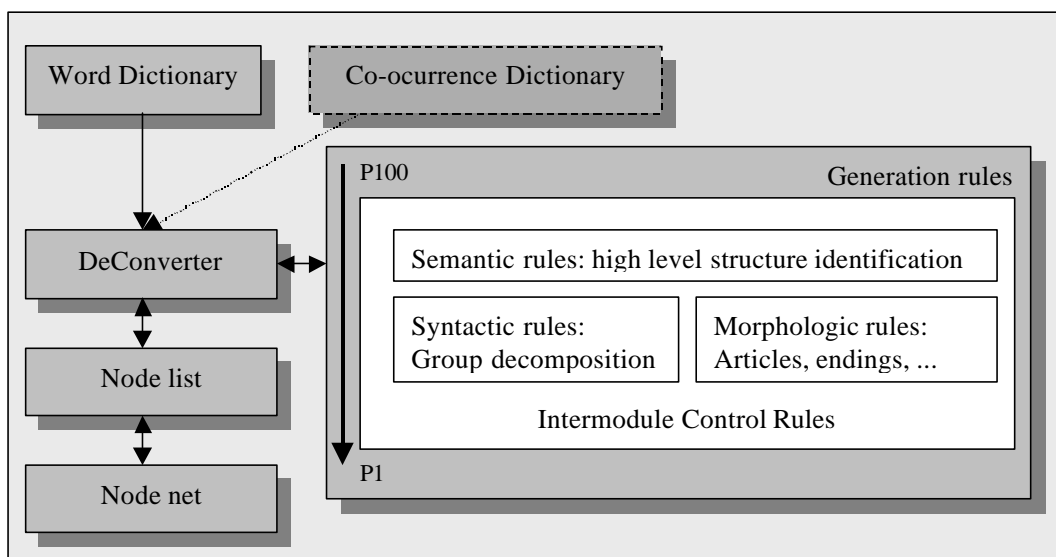


Figure 19: Integration of the rules defined with IAS Software and the rest of data structures

The main directives assumed for the development of the rules were:

At the semantic level:

- Use of verbal patterns [ASTUD-90] for the identification of the basic structure of the sentence.

At the morpho- syntactic level:

- Mapping of UNL structures in Spanish with a view to the Corporuses at our disposal, together with the grammatical reports of each linguistic unit or group of rules.
- Utilization of the same rules of control for the syntactic and morphologic modules. So that, although conceptually two passes are assumed, one for each module, the execution of a high-level rule implies a backtracking on the list in the generation

windows (“update and back” politics). The application of modularization and prioritization mechanisms permits then to avoid a double passing.

In general:

- To try to modularize and to free the rules the more possible, using priorities or imposing markers of exclusive conditions for the application of a determined group of rules.
- Utilization of the backtracking for the solving of difficult problems in the mapping of structures, as it is the only tool with a certain level of complexity.

Considering as a basis the Generator design, the different grammatical studies made and the study of the DeCo, we have carried out the implementation of the rules specific for each group identified. From that, we made a first version of the Generator of general purpose and based in the generation of simple structures.

A second approach, using the Demonstration Corpus and the previous version, led us to another additional set of rules for the solving of specific problems raised in the Demonstration Corpus. The demonstration of this last version was made in November 1997, in the Headquarters of the UNESCO in Paris.

For the new Corpus delivered on December 1997, the rule base was augmented with new general rules (See Annex 5), and the new proposed structure of the dictionary was implemented. The results of the generation can be seen in Annex 6.

Documents produced:

General comments on the UNL expressions (I, II y III) – [6], [7] and [10]

Detailed study of the UNL expressions – [8]

Functional Design of the Generator – [31]

Features of UNL Dictionary – [30]

Alternatives to the detailed design – [29]

UNL and Nominal Syntagm – [15]

Verbal Syntagm – [22]

Connection Elements – [16] and [25]

The Circumstantial Complements – [50]

Demonstration Documentation (Madrid and Paris) – [49]

Software:

Rules in the DeCo format (basic version and version amplified to demonstration) – [59]

Demonstration Corpus – [58]

Demonstration Dictionary – [57]

6.7 Rule Debugging:

As a correct debugging and validation of rules only could be made by a mechanism, the execution of test cases (corpus), the first thing to do was to evaluate the existing corpuses with a view to the last specifications [UNL-97b]. We discarded the different corpuses extracted from the text “Tower of Babel”, because they did not comply with the updated UNL specifications and because of their incompleteness in the description of the UWs. So that, following the instructions of the UNL Center, we

developed a part of the global Corpus, specifically Chapter XII of the Charter of the United Nations.

A consequence of the lack of valid corpuses was that the work of debugging was made in parallel with the last phase of development of the Demonstration Generator (based in the unique Corpus being valid).

Another point to take into account is the detection of anomalies in the software delivered by the IAS, being informed directly the UNL Center of each of these ones.

Document produced:

Analysis of Chapter XII of the Charter of the United Nations. – [55]

Anomaly report on the functioning of the DeCo – [51]

6.8 Work Environment

The activities of development were carried out on a network of PCs on the Operating System Windows NT 4.0 and using a network files system (NFS) supported by a SUN SPARC 20 Station. The tools used were mainly for tasks of documentation (Word 97), presentation (PowerPoint 97) and edition (Text Editors).

We also want to point out the importance of the electronic mail as an element of communication, not only internally between the members of the Generation working group, located in Madrid and Tokyo, but also with the UNL Center.

6.9 Glossary

<i>Co-occurrence Dictionary:</i>	Dictionary that includes relations of co-occurrence possible between two words. It is used to select the adequate word in the target language.
<i>DeCo (DeConverter):</i>	Generation System that converts the sentences expressed in UNL to the target language, through the use of a word dictionary (sometimes called “knowledge base” and a set of generation rules. Deconverter is also the name given to the language in which this generation rules are written.
<i>DicBld (Dictionary Builder):</i>	System that indexes the Word Dictionary so that it can be accessed by the DeCo.
<i>Generation rules:</i>	They are applied on the node net, carrying out different functions on this one, (such as adding a node, changing the information of a node, moving it, etc.)
<i>Generator:</i>	DeCo system with rules and a dictionary specific for the target language.
<i>Headword:</i>	Final word expressed in the target language.
<i>Node list:</i>	List composed by nodes that three-element tuples: headword, universal word and grammatical information, to which the generation rules are applied directly.
<i>Node net:</i>	Network of nodes which are Universal Words and that are the input Data of the DeCo, being the UNL expression representation the input for the Generator.

<i>SCOPE:</i>	Cluster of a part of the sub-network as if it was a unique node, so that it can be extended and processed individually as if it was another expression, that is, using a new node list.
<i>Universal Word (UW):</i>	Words that form the UNL expressions, that is the starting point of the generation. It represents the meaning of the words using an English word.
<i>UNL (Universal Networking Language):</i>	Language to exchange information on the network.
<i>Word Dictionary:</i>	This dictionary includes information on the three-element tuple: Universal Word, HeadWord and associated information. It has to be processed by the DicBld to be accessible for the DeCo.

6.10 Future And Conclusions

Broadly speaking, we can say that the evolution from the first design and ideas of the generator, to its derivation to the Demonstration Generator, presented in Paris, has gone through the development of a very basic version of the Generator that is the one from which we have to start for future expansions. This basic generator has a fundamental quality: robustness, but at the same time a problem: it is very basic. We have demonstrated that our working group, starting from the designs proposed and using UNL Center tools, is able to develop a system generating the Spanish language.

The second milestone of the project will consist in shaping each facet of the language, maintaining the robustness of the current system, so that we have a practical system and not a basic one. That is, it is necessary that the extensions that will be made cover the largest part of the Spanish language, and that they be made in a methodological and controlled way.

This extension of the generator to cover a high percentage of the Spanish language has to be carried out in the same way as until now: by the writing of grammatical reports (whether revising the existing ones or creating new ones). It is necessary to take into account that the naturalness in the generation of a language is determined in great part by the processing of apparently marginal questions of this language. We have to point out that these grammatical reports have the characteristic of making the development of the Generator independent from each sub-team of the task.

Other considerations to be taken into account for the physical implementation:

- The adequateness of the new mechanisms that could be incorporated to the DeCo.
- The study of new needs as for the new actions that could added to the DeCo (to solve new problems that could appear), as well as additional tools that could be created to improve the Generator development.
- Detailed study of the utilization of a new structuring of the Dictionary. One of the things to do would be not to catalogue the roots of the irregular forms in one unique headword, but in two (so as to choose the complete root the latest possible).

- Study and following-up of new UNL specifications that could be produced [UNL-97c].

We have a system in an embryonic state, its development consumed a great part of the design and development resources of the first year of the project. If we want to get the biggest profit out of this first year of work, we should not curtail the resources for the development during the second year.

On the other hand, the dictionary tasks, as it has more than 100.000 lexical entries, should be focused on the refining of the existing dictionary with a view to the new Knowledge base (that the UNL Center is being developing) and to the selective introduction of new terms. We must not forget that part of the success of the system lies in the quality of the Dictionary.

7. GENERAL REFERENCES

- [ASTUD-90] F. Astudillo. “*Aplicación del Método de Selección de Patrones Verbales al Español (Castellano)*”. Tesis Doctoral, Dept. Inteligencia Artificial, Universidad Politécnica de Madrid, 1990. (Ph. D. Thesis).
- [ALAR-94] E. Alarcos. “*Gramática de la Lengua Española*”. Real Academia Española, Colección Nebrija y Bello, Espasa Calpe, 1994. (Spanish Language Grammar).
- [ALLEN-87] J. Allen “*Natural Language Understanding*”. The Benjamin/Cumming Publishing Company, Inc., 1987.
- [ALLEN-95] J. Allen *Natural Language Understanding*, Benjamin Cummings 1995.
- [BOGU-88] Boguraev 1988 “*A Natural Language Toolkit: Reconciling Theory with Practice*”, paper included in [Rey88].
- [BUCH-84] Buchanan and Shortliffe eds. *Rule-based Expert Systems*, 1984.
- [COMP-92] “*Special Issue on Natural Language Generation*”, Computational Intelligence, vol. 8, num. 1, February 1992.
- [DECO-97a] “*DeConverter. Specification. Version 1.0*”, UNL Center, Institute of Advanced Studies, The United Nations University, April 1997.
- [DECO-97b] “*DeConverter. Specification. Version 1.1*”, UNL Center, Institute of Advanced Studies, The United Nations University, November 1997.
- [EWNLG-93] G. Adorni, M., Zock (Eds.). “*Trends in Natural Language Generation - An Artificial Intelligence Perspective*”. Fourth European Workshop, EWNLG '93, Selected Papers, Pisa, Italy, April 1993.
- [FERN-95] Fernández Lagunilla, Marina y Anula Rebollo, Alberto *Sintaxis y cognición. Introducción al conocimiento, el procesamiento y los déficits sintácticos*, Síntesis 1995.
- [GAZD-89] Gazdar and Mellish *Natural Language Processing in Prolog: an introduction to computational linguistics*, Addison-Wesley 1989.
- [HANSEN-94] S. L. Hansen, H. Wegener (Eds.). “*Topics in Knowledge-based NLP systems*”. Samfundslitteratur, 1994.
- [INLG-92] “*Proceedings of the Sixth International Natural Language Generation Workshop*”. Trento, 1992.
- [INLG-96] “*Proceedings of the Eight International Natural Language Generation Workshop*”. Information Technology Research Institute, University of Brighton, 1997
- [JONES-96] K. S. Jones, J. R. Galliers. “*Evaluationg Natural Language Processing Systems. An analysis and review*”. LNAI State-of-the-Art Survey, Springer, 1996.
- [LAMIQ-89] V. Lamiquiz. “*Lengua Española. Método y estructuras lingüísticas*”. Ariel Lingüística, 1989. (Spanish linguistic structures).
- [LEON-87] Leonard Bolc ed. *Natural Language Parsing Systems*, Springer 1987.
- [LYONS-97] Lyons, John *Semántica Lingüística*, Paidós 1997.
- [MATEO-84] F. Mateo, A. J. Rojo. “*El arte de conjugar en español. Diccionario de 12.000 verbos*”. Hatier, 1984. (Spanish conjugation verbs).

- [MCDON-88] McDonald, D., Bolc L. ."Natural Language Generation Systems". Springer-Verlag, New York, 1998.
- [OBER-89] K. K. Obermeier. "*Natural Language Processing technologies in Artificial Intelligence*". Ellis Horwood, 1989.
- [PASO PC 315] "*Generador de Interfaces a Bases de Datos en Lenguaje Natural*". Project sponsored by the European Community and the Spanish Ministry of Industry and Technology. 1994-95
- [PORTO-94] J. Porto, L. Gómez, y otros. "*Complementos argumentales del verbo, valores gramaticales de <<se>>, la impersonalidad gramatical, etc.*" Colección Cuadernos de Lengua Española, Arco Libros, 1994. (Several papers about minor topics in Spanish Grammar).
- [MORA-95] Morales Grela, José Ángel 1995 "*Realización de un analizador sintáctico de lenguaje natural de propósito general*". Trabajo fin de carrera presentado en la Facultad de Informática de la Universidad Politécnica de Madrid.
- [NILS-87] Nilsson *Principios de Inteligencia Artificial*, Díaz de Santos 1987.
- [PARIS-91] Paris, C. L. (ed.), Swartout, W. R. (coed.), Mann, W. C. (coed.). "*Natural language generation in artificial intelligence and computational linguistics*", Kluwer Academic, Boston, 1991.
- [RAE-96] Real Academia Española. "*Esbozo de una Nueva Gramática de la Lengua Española*". Espasa Calpe, 16^a Ed., 1996. (Draft for a new language Grammar).
- [REYLE-88] Reyle and Roher "*Introduction*" to the collection of papers that composes the volume "*Natural Language Parsing and Linguistic Theories*", Reidel 1998.
- [ROGET-97] "*Roget's Thesaurus*" – URL Reference: <http://www.thesaurus.com/>
- [SHIEB-86] Shieber. "*An introduction to unification-based approaches to grammar*", CSLI 1986.
- [WHITE-90] White and Goldsmith eds. 1990 "*Standars and Review. Manual for Certification in Knowledge Engineering*".
- [WILS-83] Wilson, Robin J. "*Introducción a la teoría de grafos*", Alianza Editorial 1983.
- [UNL-96] "*UNL: Universal Networking Language. An Electronic Language for Communication Understanding and Collaboration*". UNL Center, Institute of Advanced Studies, The United Nations University, 1996. (UNL Specification, v1.0 r1).
- [UNL-97a] "*UNL: Universal Networking Language Specification, v1.0 r2*". URL reference: <http://www.unl.ias.unu.edu>, published on April 1997.
- [UNL-97b] "*UNL: Universal Networking Language Specification, v1.1 r1*". UNL Center, Institute of Advanced Studies, The United Nations University, August 1997.
- [UNL-97c] "*UNL: Universal Networking Language Specification, v.1.1 r2*". UNL Center, Institute of Advanced Studies, The United Nations University, November 1997.