



UNL

Spanish Dictionary

- Summary of the 1st year
- Progress lines for the 2nd year
- Goals for the 2nd year
- Current state

Madrid - May 98

This brief presentation of the Spanish Enconverter offers a general outline of the whole system, and of the work already done on the different modules that make up the Enconverter.

We will finish stressing some of the benefits expected from such a system for the whole UNL project.

Summary of the 1st year

- Setting up a methodology for building the Dictionary:
 - a) close interaction with the Enconverter and Deconverter teams
 - b) selection of contents keeping the proportion between the different parts of speech and the thematic areas
 - b) use of publicly available resources



The Enconverter is a software tool designed to provide:

- a) syntactic analysis of the input sentence
- b) semantic representation of sentence's meaning
- c) automatic conversion of that representation into UNL.

The Enconverter's functionality has been divided into four modules. Each one completes a major step towards the translation of the input sentence into UNL.

The Enconverter is a software assistant for its user, who:

- a) supervises the results proposed by the sub-system, and
- b) modifies and completes those results, if necessary.

The intended user is not an NLP expert, but perhaps a person with a bachelor's degree in Linguistics or just a person with a sound training in Spanish grammar.

As an interactive tool, several usability issues should be met:

- a) efficient analysis of the input sentence and quick response to user's command.
- b) employment of terminology and representation techniques well-known in Linguistics.

Summary of the 1st year

- Design and implementation of the Dictionary:
 - a) computational independence from the other sub-systems
 - b) implementation as a relational database supported by application programs.
- Contents:
 - Over 100.000 pairs between Spanish headwords and UNL universal words.



Progress lines for the 2nd year

- Continuation and reinforcement of the established methodology
 - progress of the Dictionary according to the requirements posed by the Enconverter and the Deconverter
- Dictionary contents
 - increment in quantity and quality



We have established a set of nine basic categories (noun, adjective, verb, etc), most of them subdivided into more specific groups. For each category (and sub-category), a set of grammatical features have been defined in order to represent its grammatical properties.

The grammatical features belong to three different classes:

- a) morphological: gender, number ...
- b) syntactical: transitive, countable
- c) semantic: animate, human

At the moment, we have written rules for nominal, determinative and adjectival phrases. Adverbial, pronominal and prepositional phrases are being currently defined, and we will turn to simple and complex sentences after that.

The semantic model we plan to apply is based on:

- a) a set of primitive semantic objects (actions, entities, qualities, manners, propositions, etc.)
- b) a set of primitive semantic attributes or features that characterize these objects (animate, instrument, etc.)
- c) a set of primitive semantic relations (agent, theme, beneficiary, cause, etc.) that can be established between objects.

This representation is the base for the automatic conversion of sentence's meaning (achieved in this module) into UNL.

Progress lines for the 2nd year

- Turning to a definitive implementation of the Dictionary sub-system:

Short-term reasons against 1st year prototype:

- on-line connection with the Enconverter
- creation of visual tools for manipulating the contents

Mid-term reasons:

- interoperability with third-party applications through a standard interface (ODBC)
- performance, robustness and support offered by commercial database management systems



Dictionary: goals for the 2nd year

- Modification and enhancement of the previous design in order to:
 - facilitate the integration with the DeCo
 - automate the acquisition of new entries
- Contents' growth:
 - quantity: 100.000 new pairs
 - quality: refinement of the previous contents

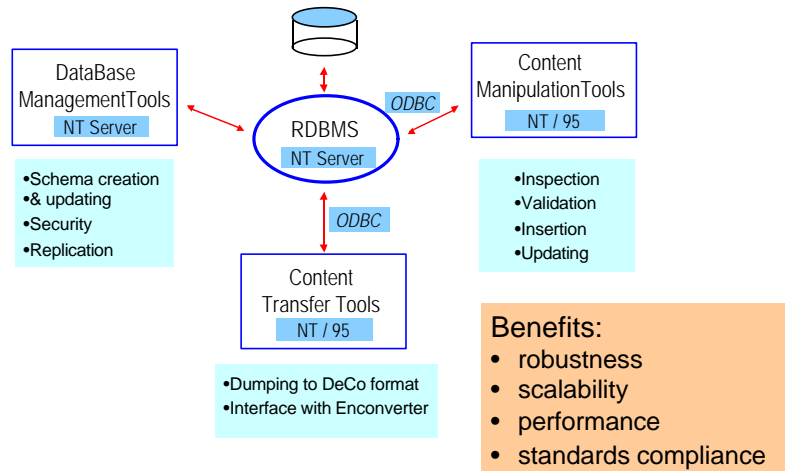


The implementation of an Enconverter along these guide-lines will have two major benefits.

Point b) is regarded as very important during the development phase for an extensive and systematic testing of the Deconverter.

Dictionary: development lines

Software architecture:



Dictionary: development lines

- Maintaining 1st year criteria regarding domain and part-of-speech balance.
- Increasing the use of publicly available resources and automatic cataloging tools
- Improving the semantic information attached to an entry applying KB's concepts.



Current state

- Definition (in close cooperation with Enconverter and Deconverter teams) of the new requirements, in particular:
 - improvements in morphology
 - definition of argument structures
 - definition of new semantic information
- Design of the new database schema
- Implementation of the new design in SQL-Server™ over Windows NT™



Current state

- Transference of the current contents to the new database
- Creation of several tools (queries, views, reports) for inspecting the data
- Classification of approximately 5.000 new entries



Lexical entries

Morphological block → Semantic block

Root	Endings
fontaner	-o
	-a
	-os
	-as

→ plumber (flapping)

Link to Uw →

Pair (Spanish form, Uw) →





UNL

Diccionario

Christèle Legéard
Roberto Jiménez
Luis Iraola

Madrid - Noviembre 98

Diccionario: cumplimiento de tareas

Tarea	Plazo	Productos
Desarrollo del interfaz de aplicaciones	27/04/98 - 24/07/98 31/7/98	Biblioteca Dinámica de Interfaz <i>Manual de referencia del Interfaz con aplicaciones</i>
Desarrollo de la herramienta de volcado	6/07/98 - 2/10/98	Programa de Volcado a formato DeCo <i>Manual de referencia del Programa de Volcado*</i>
Creación de formularios de acceso a los datos	22/06/98 - 2/10/98 16/10/98	Formularios de manejo del Diccionario <i>Documentación de los formularios*</i>
Depuración de verbos	23/03/98 - 28/08/98	<i>Informe de la depuración de verbos*</i>
Depuración del resto de categorías	23/03/98 - 28/08/98	<i>Informe de anomalías de los datos '97 y transferencia al formato '98</i>
Introducción de nuevas entradas	4/05/98 - 31/07/98 21/08/98	6.757 sustantivos y 1.846 verbos <i>Informe de la tarea 5.1</i>
Catalogación semántica	22/06/98 - 2/10/98 16/10/98	308 nuevas entradas, 789 revisadas <i>Propuesta de una ontología para el proyecto UNL</i> <i>Informe de la tarea 5.2</i>
Integración	5/19/98 – 30/10/98 4/11/98	Adaptación de los programas de volcado e interfaz con aplicaciones Correcciones al Diccionario



Diccionario: nuevos sustantivos

Periodo temporal	Sustantivos nuevos, no catalogados hasta entonces.	Sustantivos catalogados previamente, pero emparejados con distinta UW	Sustantivos ya catalogados previamente y emparejados con la misma UW
Año 97 (meses de octubre y noviembre)	3.276	1.230	1.369
Año 98 (primera quincena de junio)	3.481	1.446	0

El número total de pares <palabra española, palabra universal>
asciende a: **20.284**



Diccionario: nuevos verbos

Verbos catalogados	Nuevas palabras universales	Entradas léxicas	Pares: <forma verbal, UW>
1.846	1.503	2.762	179.530

El número total de pares (sustantivos y verbos) es de: **199.814**



Diccionario: cifras globales 97+98

- Entradas en forma de cita: 15.900
 - Sustantivos: 11.558
 - Verbos: 2.414
 - Adjetivos: 1.738
 - Adverbios: 272
- Palabras universales básicas: 18.331
- Palabras universales: 21.771
- Entradas léxicas completas (enlaces Hw-Uw): 23.912

