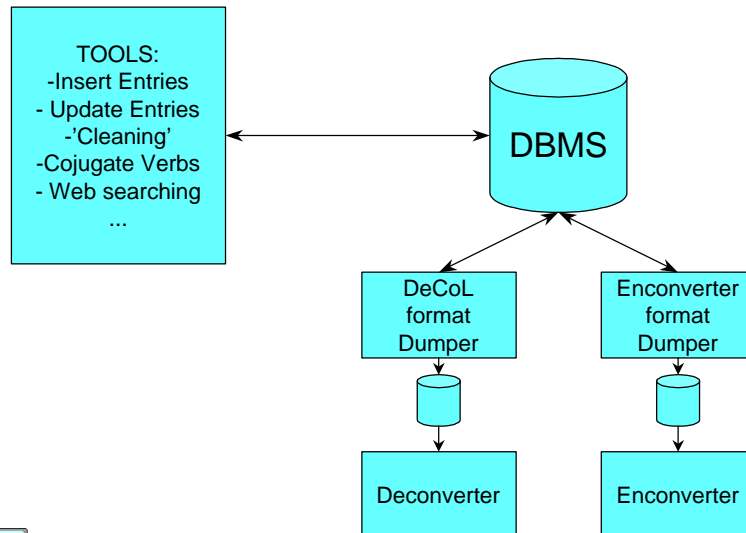


UNL Dictionary

Spanish dictionary for UNL

Notes on the Spanish Dictionary System.

Architecture



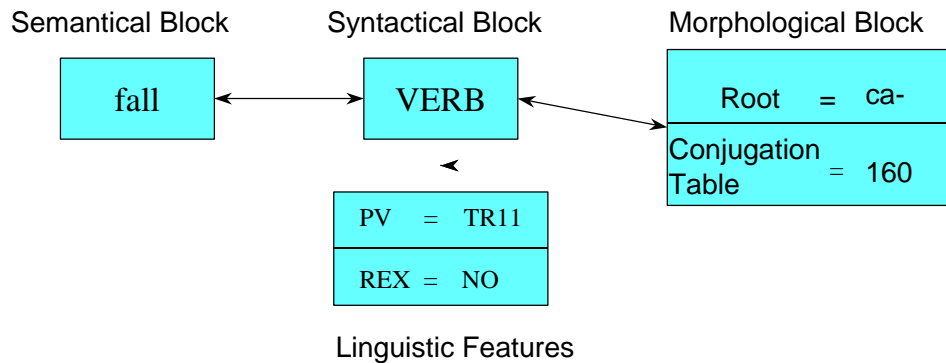
This slide shows the whole system architecture for the Spanish Dictionary.

All the dictionary information is stored using a distributed relational DBMS (Database Managing System). The DBMS runs in a computer which works as a server, and is exclusively used for this task. There are other computers connected to the internal network which may access the dictionary information as clients through specifically developed client applications.

The main tools developed to the point are the following:

- **Dictionary maintenance tools.** They allow inserting and updating Universal Words into the dictionary with all the necessary linguistic information and Spanish morphology of the word. They also allow simple queries by Universal Word or Spanish word.
- **UNL DeCoL format dumper.** Dumps all the information stored in the relational database into the Latin Deconverter's format.
- **'Cleaning' Tool.** Removes any residual information from the relational database.
- **Statistics Tool.** Extracts quantitative information about the contents of the dictionary.
- **WEB Querying Tool.** Allows queries both by Universal Word or Spanish word through a Web browser.

Internal Structure of the Dictionary



The internal structure of the dictionary (relational format), is shown in this slide.

The universal words are stored in independent blocks called Semantical Blocks. Each of these has one or many Syntactical Blocks linked to it. These blocks store the different syntactic categories (verb, adjective, ...) which may correspond to that concept in Spanish. All the linguistic information related to each word is represented by many Linguistic Features linked to the Syntactical Block.

Finally, the morphological information of the different Spanish words which may correspond to the Universal Word, is stored in Morphological Blocks linked to each Syntactical Block.

There is an additional kind of block which is related to those Spanish words (Morphological Blocks) describing Spanish verbs. As in Spanish language, verb endings change following 90 different patterns, this information is stored in separate tables related to the Morphological Blocks that, in this specific case, only contain information about the root of the verb.

Dictionary Contents

- **Pairs:** **101,904**
- Universal Words: 9,807
- Morphological Blocks: 11,328
 - Sustantives: 6,862 blocks
 - Adjectives: 2,006 blocks
 - Verbs: 1,840 blocks
 - Other categories: 620 blocks.
- We have made a **carefull selection of entries** according to:
 - Use of a thematic tree to select the entries.
 - Keeping a balanced number of words from each category.



The contents of the dictionary have been carefully selected so that they are leveled both in the amount of words from the different syntactical categories and in the amount of concepts which appear in different kinds of texts.

At this point, our dictionary contains 101,904 pairs, that is to say 101,904 correspondances between Universal Words and different Spanish inflected forms.

As you can see, there are 9,807 different concepts (Universal Words) stored in the dictionary. These concepts are associated with 11,328 Morphological Blocks, which are basic words in Spanish language. Each of these words generates different forms depending on number, gender, ...

Example of Word Introduction

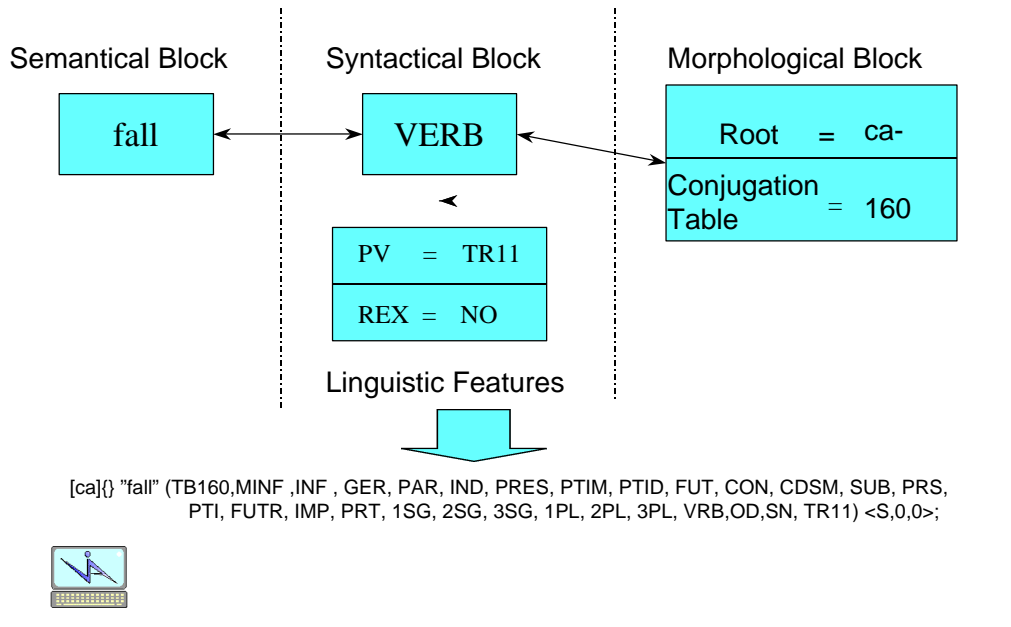
Conectando a la base de datos...HECHO Recuperando atributos...HECHO	1	Raiz: IAS Tabla de conjugación [0]: Morfema Masculino: Morfema Femenino: Morfema Plural (masc): Morfema Plural (fem):	5
[* Mantenimiento del diccionario *] 1.Nueva entrada al diccionario 2.Buscar UW 3.Buscar Palabra 4.Numero de entradas 0.Salir Opción>1	2		
UW? IAS(icl>organization) Buscando ...	3		
Categoría: SUS introducir rasgos (s/n)? s Atributo: CAT2 Valor:PRP Atributo: GEN Valor:MAS Atributo:	4		

1. DB Connection
2. Main Menu
3. UW introduction
4. Features intro
5. Morphological intro

In this slide we can see the execution of one of the Dictionary maintenance Tools:

- First of all, the program connects with the remote database.
- Once in the main menu, option 1 allows entering new UW's while option 2 lets you modify the existing ones.
- Introduction of the UW (IAS(icl>organization)) and database search to check that it was not already in the database.
- Introduction of linguistic features:
 - Category: Sustantive (SUS)
 - Subcategory: PRP
 - Gender: Masculine (MAS)
- Morphological introduction: IAS has no gender-number variation, so 'root' is the only field to be filled.

Example of DeCoL format.



This slide shows an example of the Transf execution.

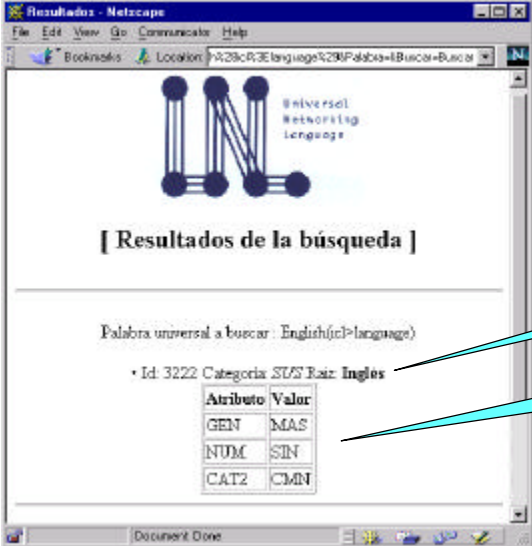
The internal representation in the dictionary has one semantic block (fall), one syntactic block (verb category), linguistic features (verbal pattern TR11, etc.) and one morphological block (root: ca- and a conjugation table '160').

The transfer process forms 1 headword with a link to the conjugation table.

Following the links, the application selects the syntactic block and dumps their information (VRB) and their associated linguistic features (PV and REX).

The last step is to assign the Universal Word to each entry previously formed.

Web Access Example



The screenshot shows a Netscape browser window with the title 'Resultados - Netscape'. The address bar contains the URL 'Location: %28ic%2Elanguage%28Palabra=Buscar+Buscar'. The page content includes a logo with the text 'Universal Networking Language' and a heading '[Resultados de la búsqueda]'. Below the heading, it says 'Palabra universal a buscar: English(icl>language)'. A search result is shown with the text '* Id: 3222 Categoría: S/S Raíz: Ingles'. Below this, there is a table of linguistic features:

Atributo	Valor
GEN	MAS
NUM	SIN
CAT2	CMN

Two callout boxes point to the search result text: 'Root (morph)' points to 'Raíz: Ingles' and 'Linguistic features' points to the table.

In this slide we can see the search results.

The page contains the original word to search (English(icl>human))

The next line shows the semantic block identifier, the syntactic category and the morphological root.

Next, we can see a table containing all the linguistic features corresponding to this syntactic block.

If more than one syntactic blocks are found, they will be represented by a set of tables following the first one. If no blocks are found, an error message appears.

Next steps

- Further compression of morphology
- Improve performance and reliability of DB engine.
- Increase the number of entries and information associated to them
- Content validation



Main tasks to be carried out in the near future:

Improving access time, store capacity, and reliability of the Database.

The programs must be enhanced to make them easier to handle. Another thing we are considering, is allowing semi-automatic word introduction.

Finally, all the contents of the dictionary must be validated to assure the quality of every entry.

The transference to the IAS DeCoL format must be changed in order to achieve further compression of the Spanish morphology.