An environment for management and extraction of taxonomies from on-line dictionaries

Alicia Ageno
Irene Castellón
Germán Rigau
Mariona Taulé
Sivard Cardoze
María Antonia Martí
Horacio Rodríguez
Felisa Verdejo

Report LSI–91–25

AN ENVIRONMENT FOR MANAGEMENT AND EXTRACTION OF TAXONOMIES FROM ON-LINE DICTIONARIES

Alicia Ageno (1)
Sivard Cardoze (2)
Irene Castellón (1)
M. A. Martí (3)
German Rigau (1)
Horacio Rodriguez (1)
Mariona Taulé (3)
Felisa Verdejo (1)

(1) Universitat Politècnica de Catalunya
(2) Katholieke Universiteit Brabant
(2) Universitat de Barcelona

June 1991
LSI Departament
Universitat Politècnica de Catalunya
Barcelona
SPAIN

ESPRIT BRA-3030 ACQUILEX - WP NO. 020

An environment for management and extraction of taxonomies from on-line dictionaries

1. Introduction.

Three main problems are concerned in using Machine Readable Dictionaries (MRD) as a source of lexical and conceptual knowledge for Natural Language Processing Systems.

- How to extract the information contained, in a rather unstructured form, within the MRD entries.

- How to represent this information in a way theoretically sound allowing a concise and efficient implementation while offering an easy form of navigation, manual modification and updating.

- How to exploit the data acquired and represented, for NLP applications.

All three problems are faced within the Acquilex project. This paper tackles with the first one, the extraction of semantic information, basically taxonomic relations, in a (semi) automatic way, from a Lexical Data Base (LDB). This LDB structures and organizes in an accesible way all the information contained in a Machine Readable Dictionary (MRD).

The MRD on which the development is based, is the Spanish Monolingual Dictionary Vox [Vox 87].

Our approach assumes an interactive process for extracting taxonomies, as well as other semantic relations, from the Vox dictionary.

The organization of this paper is as follows: after this introduction, section 2 presents an overview of the environment, its design principles and capabilities. Section 3 is devoted to the tasks involved in extracting taxonomies from the Vox dictionary. In this section a detailed example is presented. Section 4 explains the tasks concerned with heuristics evaluation and management. Section 5, to sum up, presents some conclusions and outlines the current state of development.

2. Overview of the system.

Taking into account the need of having available a working prototype to be used for extracting taxonomies (within the temporal scope of the project) the first requirement was to build the prototype as soon as possible. This constraint has imposed severe restrictions on the system design and implementation.

The main considerations for the design have been:

- Extracting semantic information from dictionary entries states a problem that cannot be solved in a fully automatic way. In some extent, the choices made by the system must be validated and confirmed by a human expert. This implies the use of an interactive environment for performing such a task.

• Another consideration taken into account is the reusability of the resulting data structures for other environments, especially the convertion process to the LKB [Ageno et al 91b][Ageno et al. 91c].

• A great level of reuse of both tools and methodology from others partners, that is, the use of Cambridge LDB software [John Caroll 90] (including Alshawi's FPar parser [Alshawi 89, 90] and SanFilippo's Seg-Word morphological analyser [SanFilippo 90a, 90b] as well as Copestake's approach for extracting taxonomic information from dictionary definitions [Copestake 90a] [Copestake 90b].

• A realistic consideration about the tasks to be accomplished and the knowledge involved, states the need of a flexible system where, initially, a great amount of human intervention would be required and later, the autonomy of the system would be improved incrementally. However, human intervention will nearly always be necessary to validate the decisions automatically taken.

• Some of the tasks involved in the extraction of semantic information (for instance, the analysis of dictionary definitions) are very time-consuming and, thus, cannot be integrated in an interactive process. So, the system must allow the cooperative performance of both interactive and batch processes.

2.1 The Source.

The main source for semantic relations extraction (first of all taxonomy extraction) is the LDB containing the Vox dictionary.The functionality of the LDB is described in [Caroll 90].There are obvious reasons for chosing the LDB rather than pure MRD´s as sources of information.The loading process of the Vox dictionary [Castellón et al. 90] [Castellón et al. 91] permits the conversion of MRD entries, like the following:

> [EP[j2]I) cacho [k1](l. [k2]calculu, [k1]piedrecita) [k2]m.
> [k1]fam. Pedazo pequeño de alguna cosa.[k2] 2 [k1]Cierto juego de naipes.[k2] 3 Méj. [k1]y[k2] P. Rico. [k1]Participación pequeña en un número de la lotería.[EP[j3] [j6]Sin.[j7] [k2]1 [k3][k1]v[k3].
> Pedazo.

that appears in printed version:

> **I) cacho** (*l. calculu , piedrecita*) m. *fam*. Pedazo pequeño de alguna cosa. *2* m. Cierto juego de naipes. *3* m. *Méj. y P.Rico*. Participación pequeña en un número de la lotería. SIN. 1 v.Pedazo.

to the following lispified entry:

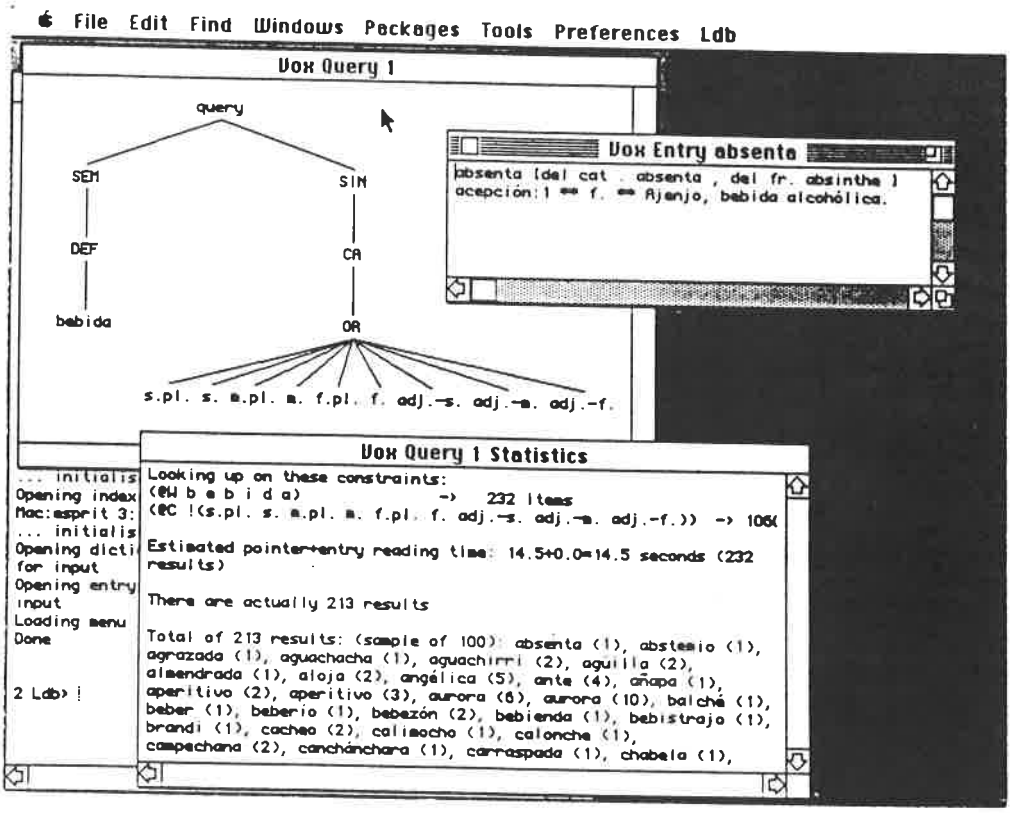
```
((cacho )
(NH I)
(ETIM l. calculu , piedrecita )
(Sense 1)
(CA m.)
(REG fam.)
(DEF Pedazo pequeño de alguna cosa.)
(Sense 2)
(CA m.)
(DEF Cierto juego de naipes.)
(Sense 3)
(CA m.)
(GEO Méj. y P.Rico.)
(DEF Participación pequeña en un número de la lotería.)
```

(RELA 1)
(TIPOR Sin.)
(TXR 1 v.Pedazo.)
)

The set of all the lispified entries is the input of the LDB system. Each entry is indexed in the LDB by a set of fields allowing an efficient access.

Obviously, the definition field is the most important one. The definitions of the Vox dictionary are written in plain text. There is no limited vocabulary, such as is the case with the LDOCE. Furthermore, the forms of the Spanish words vary much more than in English. Therefore, we only index the first 10 non-functional words in each definition. Functional words are determined by means of a statistical extraction process, which extracts the first 50.000 distinct forms appearing in the Vox dictionary. Those which appear the most frequent, are considered as being functional words. From this list, we can eliminate those which may be relevant for a later analysis, such as the first appearances of: "lo", "que", etc in order to be able to find the definitions in which the generic term is formed by words which should be considered as being functionals.The following example shows one of these cases:

Substancia I (1) : Lo que hay de permanente en un ser.



Window 0.

Using the LDB, we can perform all kinds of queries to find, for example, the senses in the Vox dictionary with the morphologic category noun which have a particular word in their definitions.

As we can see in Window 0, we have asked the LDB system for all those dictionary entries which have a definition of the category noun and containing the word "bebida" (drink). Some of these senses will have "bebida" as a generic. In this way, "absenta" ISA "bebida".

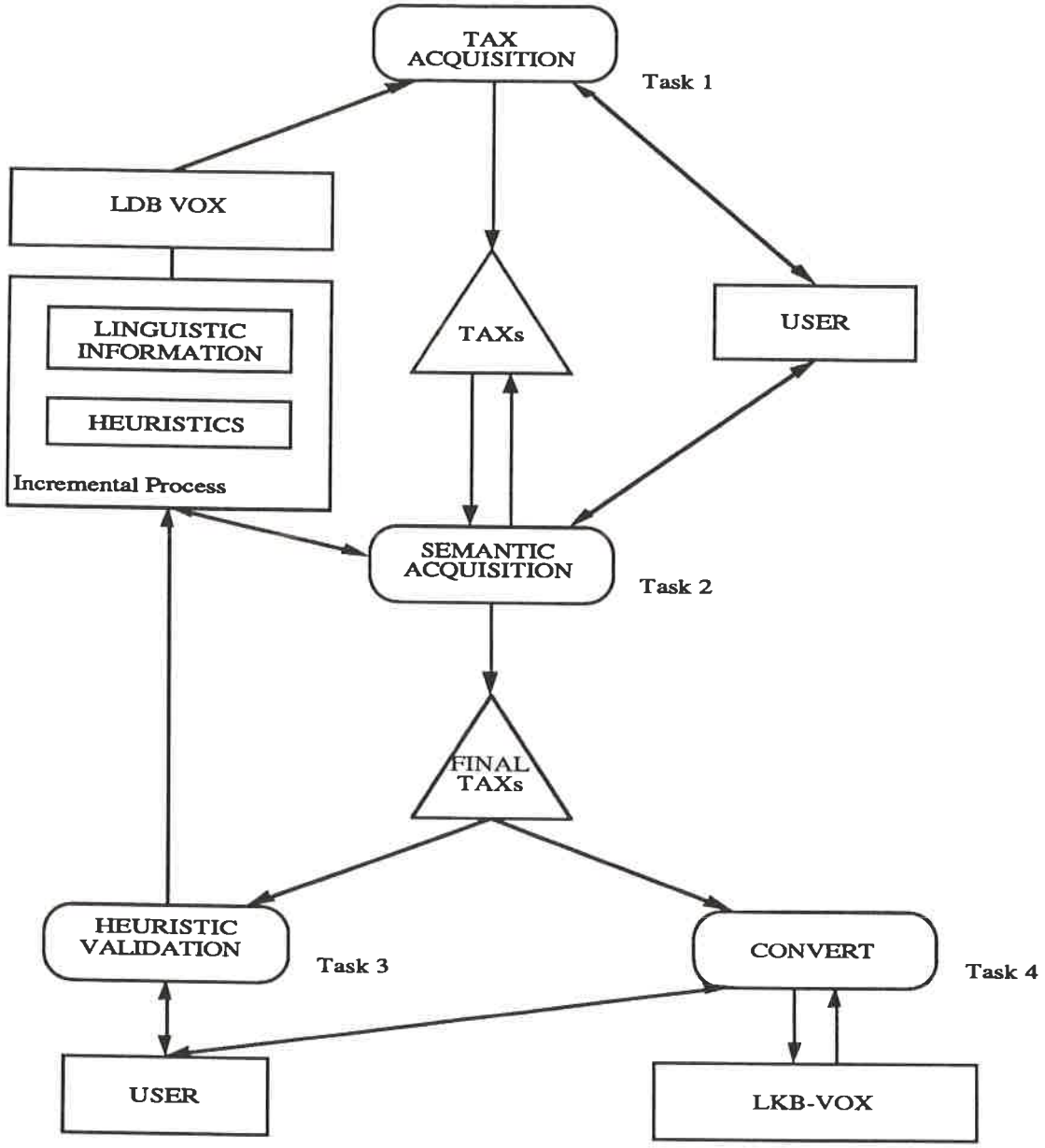


Fig. 1: General Scheme of the System.

The system operates under two modes: an acquisition mode and a validation mode. In the acquisition mode, the system facilitates the interactive construction of the taxonomies followed by the semantic analysis of the definitions contained in them. The semantic analysis of taxonomy nodes needs no user intervention and can be made by a batch process. On the validation mode, the aforementioned analysis are validated and corrected by means of an interactive process. The user, then, must decide either to modify the grammar and then redo the semantic analysis with a different grammar or correct the outcome by hand. This process may repeat itself as many times as the user deems necessary. Therefore, the grammar of each taxonomy is modified in an incremental way, until achieving an optimal result. Next, the resulting taxonomic structure (we assume that this structure is always correct) can be confronted with different sets of heuristics to improve them. Once again this operation does not need human intervention. The results are, however, also taken into account.

To facilitate the clarification of the figure, the extraction process of non-taxonomical semantic relations has been placed prior to the heuristics validation process, but both are independent of one another, can be carried out in whatever order and run with

The LDB is a static source. Other sources of knowledge for the extraction process, improved in an incremental way, include:

• The sets of rules for morphological and semantic parsing used by Seg-Word and FPar respectively.

• The LEXICON for the words which do not appear in the dictionary and those which are frequently consulted by Seg-Word. The contents of LEXICON are records containing a word and a list of possible morphological categories.

• The set of heuristics to be applied in different choice points.

All these knowledge sources are dynamic. At first, the system doesn't cover all the possible cases. As we construct taxonomies, we must incorporate to both the set of morphological rules as the LEXICON, those new cases which appear and that we wish to take into consideration. Furthermore, we must develop different syntactic-semantic grammars, according to the thematic environment to which the taxonomy pertains [Ageno et al. 91a]. In the beginning, these grammars will also not extract all the information that we wish. We must improve them gradually until achieving the desired results.

2.2 The Process.

Our system carries out four differents tasks: taxonomy construction, semantic relations extraction, heuristics validation and knowledge integration into the LKB, as shown in figure 1. The first one consists in the extraction of the taxonomy structure which lays underneath the Vox definitions, starting from a top entry. These top entries, can be easily located by their high frequency of appearance as a genus of the definitions [Copestake, 90b]. The second, the extraction of the other semantic relations which appear in the definitions of the taxonomy already created. The third task, the heuristics applied in the taxonomy construction are validated. Finally, all the information acquired is integrated in the LKB.

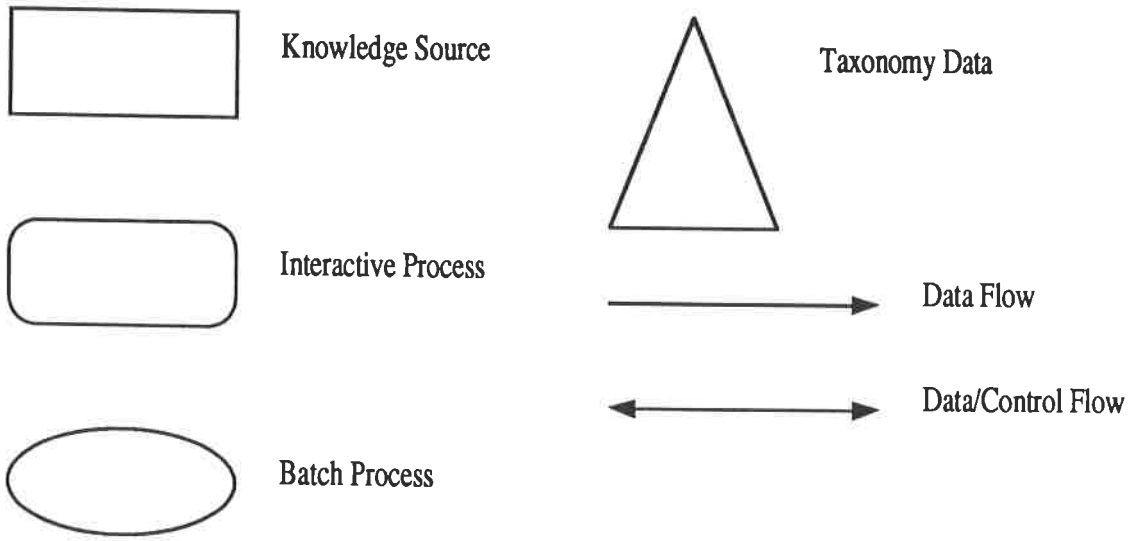


Fig. 0: Description of the symbols.

incomplete taxonomies.

The linguistic knowledge must be acquired incrementally as different taxonomies are being constructed as we go along, since we do not dispose for the moment of complete analysers of definitions for the Vox dictionary.

All the available tools had to be improved so that they could adapt themselves to the new cases which were going to be studied. Thus, performance of Seg-Word must be improved so that it is able to analyse new cases which have not occurred up till now. The words which cannot be analysed by Seg-Word must be added to LEXICON, or well the rules of Seg-Word must be modified. New cases must be taken into account in the posterior optimization process.The generic extraction grammar must be adapted to each definitions typology that is to be treated. Similarly, different grammars must be created for each taxonomy we wish to study.

As a last remark, our Tax-Build system is fully compatible with the LDB system, so that at any point of the process of construction, the user can make as many, and as complex queries to the LDB as he deems necessary .

2.2.1 Taxonomy Acquisition Task.

This module is in charge of the extraction of the taxonomies which lay under the definitions of the Vox dictionary [Amsler 81]. However, the parsing process to obtain the ISA relation from the definitions, is the same one which extracts the other semantic relations.

Figure 2 illustrates the process of extracting taxonomic relations must solve two principal problems: the extraction and the disambiguation of the generic term [Copestake 90a].

In our case, the problem of the extraction of the generic term is solved by means of the FPar syntactic-semantic analyser [Alshawi 89], with some special grammars for the extraction of generic terms. Given a sense, using this parser, we can detect its hyperonyms [Ageno et al. 91a] as well as other semantic relations.

The input of the Alshawi analyser, is a sense augmented with its morfological analysis. The morfological analysis is carried out using an, optimized, Seg-Word [SanFilippo 90a] analyser (see section 3.2.2).

Genus extraction. Given a Top entry, provided by the user, a search for all its occurrences within the Vox senses definitions, using the LDB, is performed. In some of these ocurrences the Top entry will not be the generic term of the definition. The user is assisted in his decision by means of the output of the FPar analysis. This result, might be correct or incorrect. In both cases it is noted down for a posterior study . If the word Top was indeed the generic term, then the disambiguation process proceeds. If the contrary holds true, the next occurrence is taken into account.

Genus disambiguation. As a Top entry might have more than one sense, once a sense is located as hyponym of the Top entry, it has to be linked to one of its senses. This process is also assisted by the program by means of a set of heuristics which determine which sense of the Top entry has more probabilities of being the hyperonym searched for. The program only suggests which might be the hyperonym, the user must ratify or rectify the system's proposal. Only when the Top entry has a unique sense the assignation is done automatically.

Whenever new senses are examined to be included in the taxonomy, and after the set of heuristics is applied and the result is obtained, the user must confirm the choice

proposed by the system (in fact a set of meta-heuristics is in charge of this task). The decision of the user can either confirm or override the system's proposal (i.e., the sense is included in the taxonomy or not). Then the success or failure of the heuristics applied of such heuristics is stored for further evaluation.

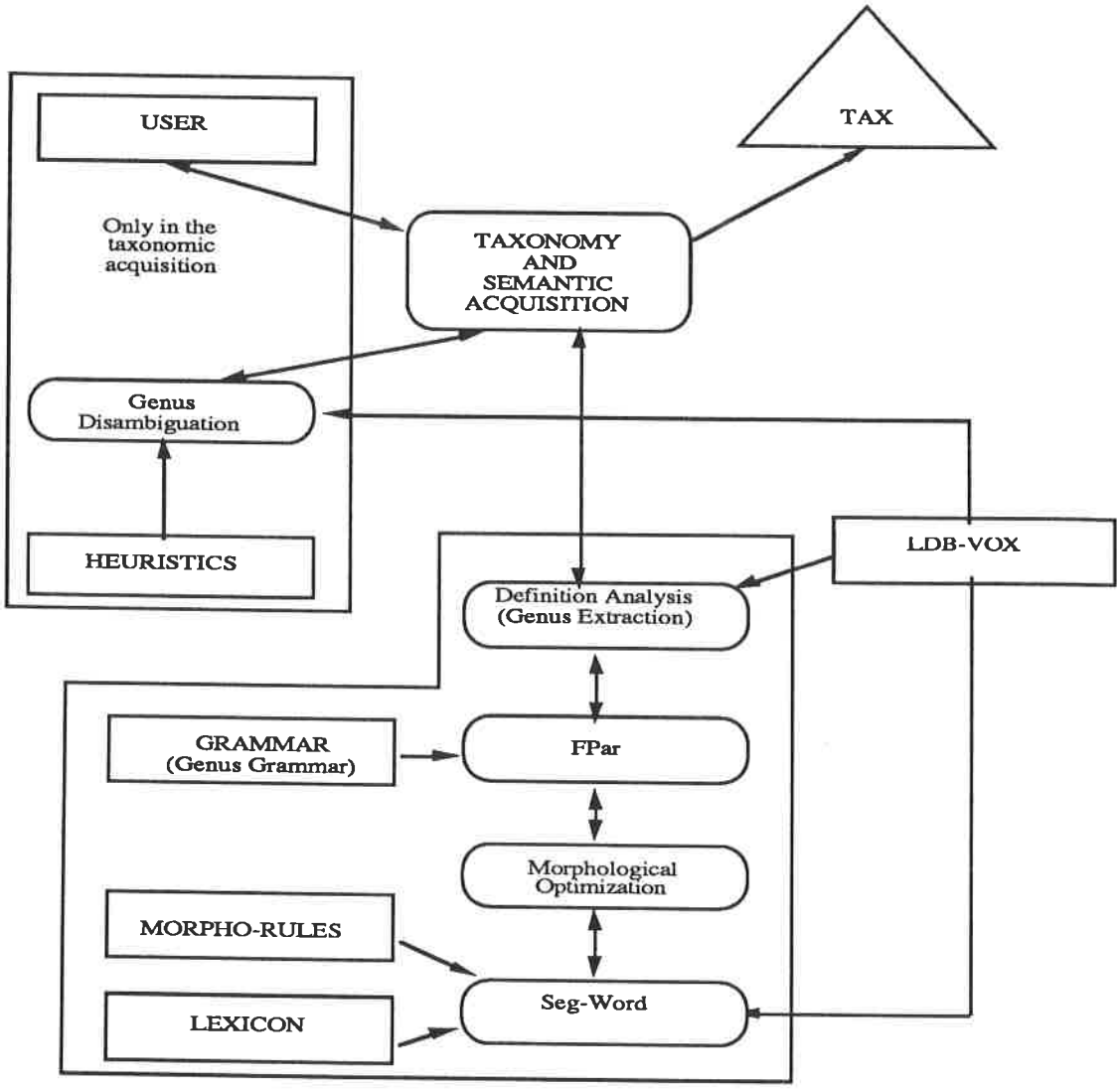


Fig.2: Acquisition.

We have prefered to express disambiguation criteria as a set of parametrized heuristics. The system is provided with capabilities both to test the accuracy of the heuristics set and to improve it allowing deletion, addition or modification of any of them. The modification of the relative weight of each heuristic is also a useful capability for tuning the system.

As the taxonomy is being constructed, some information already related to the nodes of the taxonomy is inherited by the current node. In this way, the heuristics work not only with the information available in the disambiguation node, but also with the information already acquired from the parent node(-s).

We will discuss the structure and procedures for these heuristics in section 4, as well as the evaluation process of such heuristics.

When a sense is defined as a hyponym of another, the entry to which the sense

belongs, becomes the next Top entry. If a sense does not generate any hyponym, it is converted into a terminal node of the taxonomy .

The realization of a complete taxonomy is a long process, due to the time spent on parsing the entry definitions, but the system provides the possibility to perform this process in an incremental way. The system also allows further modification of the taxonomies, either to eliminate or to redo parts of them.

2.2.2 Semantic Acquisition Task.

Once a taxonomy is created, the treelike structure is available in which all the senses included are connected with their hyperonym (except for the first Top entry) and their hyponym (except the terminal senses).

The next step (semantic acquisition) consists in realizing a similar process to the taxonomy building one, but with a different grammar and without user intervention. This batch process is called definition analysis (see figure 3). The grammar, of course, must be more complete and complex than the one for generic term extraction, because it must allow the extraction of the 'differentia' [Calzolari 90][Ageno et al. 91a] from the definitions associated to the nodes of the taxonomy. Different grammars are created for each taxonomy being dealt with, as, for instance, the information which can be extracted from "persona" taxonomy is different to the one related to the taxonomy of "substancia". The result of this parsing is an analysed taxonomy.

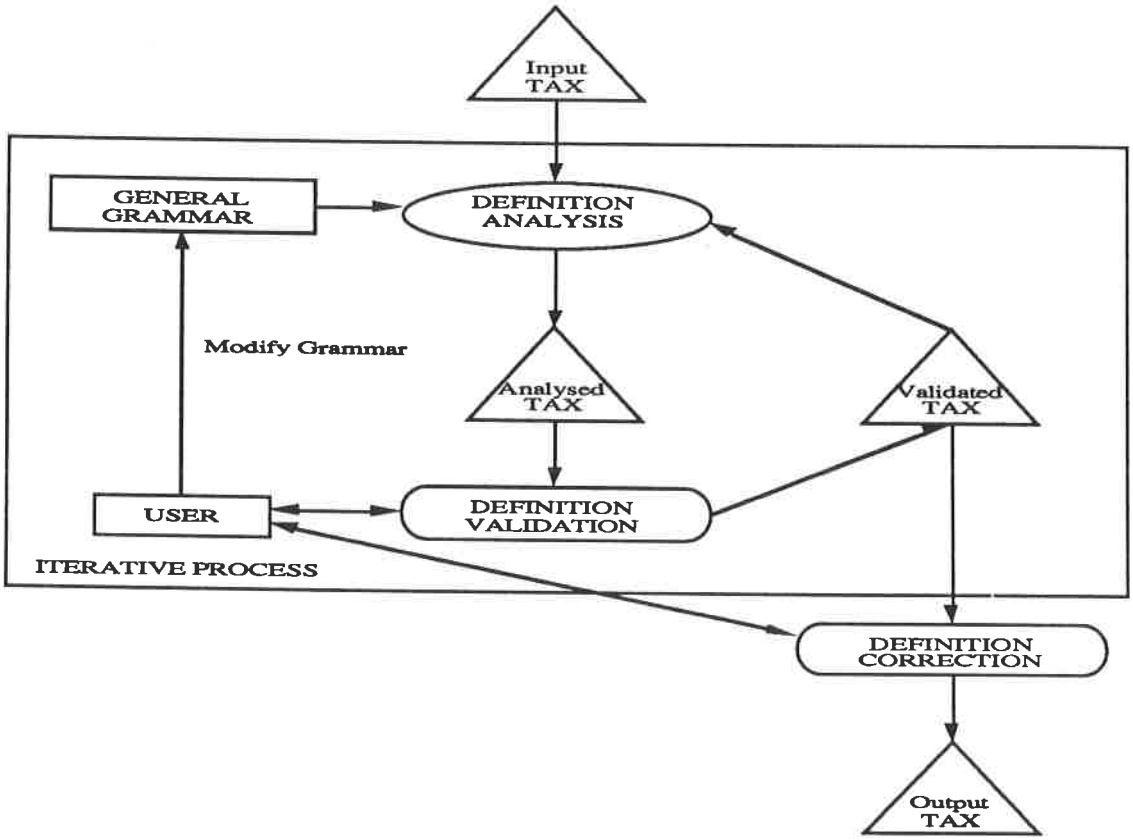


Fig. 3: Semantic Acquisition.

Since the extraction process of this information can give erroneous and/or incomplete results, a validation of the output is necessary .This process is interactive and permits one to observe the increase of the extracted information as different analysis are being carried out. Hence, any taxonomy can contain multiple analysis. Viewing the result of the different analysis, the user can determine the modifications to be performed in the

final grammar and/or the morphological module. This process can be carried out as many times as desired.

Finally, when the user either agrees on the result or he considers that it cannot be improved, the process of correction of the definitions analysis must be performed. This process is also interactive and permits the user to correct, if necessary, and to validate each one of the analysed definitions of the taxonomy. When this process terminates, the taxonomy is also ready for the interactive creation of new lexical entries for the LKB starting from the validated analysis. This is done by means of the convertion program [Ageno et al. 91b,91c].

2.2.3. Heuristic Validation Task.

After a taxonomy has been created, the heuristics applied for its construction are checked automatically to obtain the values of success and failure of each one. Then, with this information, the user can adjust the parameters of the heuristics involved, in order to improve their performance.

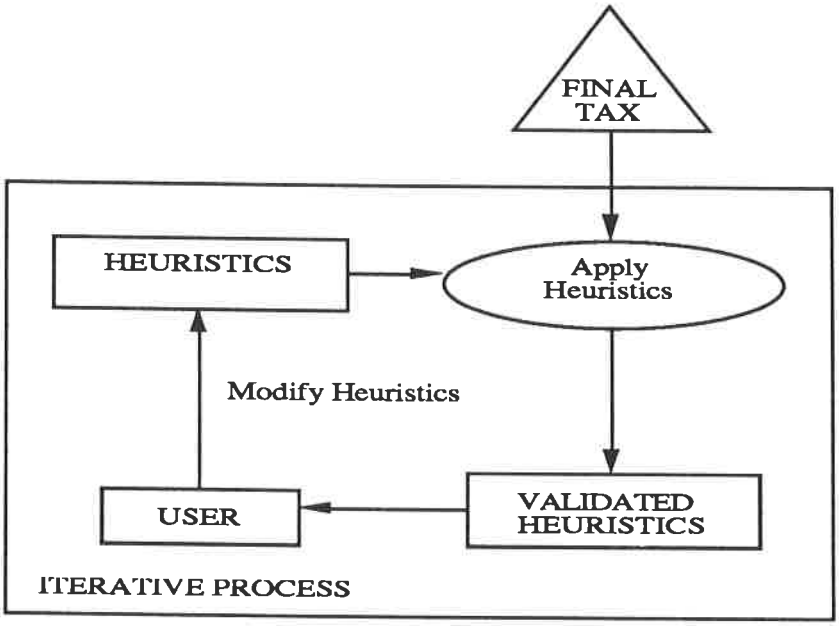


Fig. 4: Heuristic Validation.

Furthermore, new heuristics can be contrasted and evaluated using the taxonomy already constructed.

Some of the heuristics applied for sense disambiguation are described in section 4.

3. Extraction of semantic relations from Vox dictionary

The usual methodology for building a LKB from an MRD follows three major steps:

- Treatement of the MRD in order to include its information in a LDB structure.

- Specification and implementation of computational and hand-handled methods to extract implicit and explicit information contained in the LDB fields.

- Definition and construction of the LKP from the results of the 2nd step.

Among the different kinds of relations that can be extracted from dictionary entries in

order to link the different senses, taxonomic relations play a central role.

From the point of view of dictionary definitions, a special taxonomic relation, the ISA or hyperonym relation, can be established via the genus term of the entries. Other relations can be found in the different parts of the definitions. The basic source of knowledge for extracting other semantic properties from dictionary senses is the analysis of the definitions. The main tool to carry out this task is the Alshawi parser. Although all kind of semantic properties, namely ISA relation, other taxonomic relations and other descriptive properties can be extracted by this method, we have paid special attention to the extraction of the hyperonym relation. The other taxonomic relations are regarded as complementary results for now.

Next, we will focus on the taxonomy construction from a LDB containing the Spanish Vox dictionary .

We assume that the unit in the conceptual representation is the sense (there is not a one-to-one mapping between dictionary entry senses and these conceptual units, as we will see later).

The main task in the extraction process will be to attach semantic information to these conceptual units.

We can distinguish two kinds of semantic information to be extracted from LDB. Some semantic properties must be encapsulated into the conceptual unit and can be seen as the internal structure of such units. Other semantic properties links any conceptual unit to others related with.

These links constitute the structure of the knowledge base. Among them, taxonomic relations, and specially the ISA relation, play a central role.

In order to build a taxonomy from the Spanish Vox dictionary we will consider the following information presently available, in the LDB and susceptible of being used during the extraction process:

- the entry;
- the part of speech;
- the definition;
- particular uses: Fig., not usual, informal, etc.
- the subject matter: biology, medicine, etc.
- geographic information: Amer., Arag., etc.
- semantic relations: synonyms, antonyms.

In the case of IS-A relation, the main source of information is the definition. The use of the other kind of information mentioned above is restricted to the definition of some of the heuristics.

For extracting other semantic relations, the definition would be the main source of information. Some other information available would also be useful too, e.g. the semantic relations appearing explicitely in the entries, as RELA tags (synonyms, antonyms, ...).

Our approach to derive taxonomic relations from the Vox dictionary has been decided, taken into account a set of important restrictions, derived from the characteristics of our source dictionary and the lack of lexicographic material useful for automatic processing of the MRD environment. These restrictions increase the need of human intervention during the process.

The extraction of information implies selecting among different possibilities in several

points (for instance, genus extraction, sense disambiguation, stopping or pruning a branch of the taxonomy, etc...). The lack of lex. ... for automatising these entries, increase the need of human intervention.

3.1.- Restrictions.

The main restrictions due to the characteristics of the Vox dictionary and are considered below, in section 3.1.2.

Other restrictions refer to the lack of sufficient lexicographic material available. Specifically, we have limitations on analyzing morphologically all the forms (potentially) appearing in text definitions and on parsing all these definitions. Section 3.2.2. elaborates on this problem.

3.1.1.- Some comments about the source dictionary

In general, the information contained in a MRD appears in a raw form. The loading of this information on a LDB implies a more structured form of organizing this information allowing an easier way of accessing to it, but it doesn't mean in any case the addition of new information related to some kind of recodification.

This mapping process involves usually only minor transformations of the information stored in the MRD.

The power of the LDB depends on the amount of information contained in the source dictionary, but also on the form this information is expressed.

Some dictionaries, especially the LDOCE, with a large background in lexicographic studies, present this information strictly organized and codified. This information is, then, easily structured and loaded into the LDB, allowing the use of very precise procedures for extracting semantic information.

Unfortunately, this is not the case with the Vox. There are several features that prevent the application of the approach used in the LDOCE:

- there is no restricted vocabulary in the definitions;
- there are no semantic codes ;
- there are no box-codes;
- the explicit appearance of semantic informations like geographical, thematic, etc. indications appear in a non-systematic way.

These features are crucial, for instance, for handling the disambiguation of word senses with respect to a given genus term and while building a taxonomy in Copestake´s proposal [Copestake 90a,90b].

3.1.2. Morphological analysis.

Another important restriction arrives from the necessity of full coverage of morphological analysis of the words contained in dictionary definitions. This need is imposed by the use of a parser in order to extract the genus and 'differentia' from definitions.However, we do not need the morphological attributes (number, gender etc...) , only the list of POS associated to every form which appears.

Three main options were considered when tackling this problem:

- The first approach was to carry out a complete disambiguated morphological

analysis of all the definitions, i.e. not only analyse words but also analyse them in their contexts. This meant obtaining a list of pairs <word, pos> for each definition, which involved the use of a morphological analyzer with (micro) syntactic disambiguating capabilities. The Spanish analyzer from Pisa [Ratti et al. 80] seemed to be the most convenient tool incorporating all these facilities and was therefore considered. Unfortunately, the coverage would not have been complete due to, among other reasons, the great amount of americanisms and dialectalisms present in our dictionary and the published success ratio of the Pisa analyser (about 80 %).

- The second approach involved a full analysis of words appearing in the dictionary definitions, but now in an isolated way. In this case the output for each word would have been simply a list of possible POS. Of course the same Pisa analyser could be used, as well as the SIPA morphological analyzer of Spanish developed by M.A. Martí [Martí 85, 88]. But the problem with the latter was also the degree of coverage: the SIPA analyser was created to cover an economic and political sublanguage and it would have been highly expensive for us to customize it for a wider corpus.

- A third approach involved the use of dictionary entries as the source for word categorization. We could follow two possible directions: incorporate this information to a separate lexicon and, so, build it in an incremental way, or consult online this information without storing it in a separate lexicon. The first possibility allowed some kind of human intervention related to discarding some of the categorizations existing in the LDB. Both possibilities implied the use of Seg-word[Sanfilippo 90a].

Finally, we have chosen an enhanced version of the third option, for, on one hand, the other ones presented the above mentioned important shortcomings, and, on the other hand, this one seemed to be the most consistent choice with the methodology we intend to follow :

- Reuse of tools from other partners :

  - SanFilippo's Seg-Word morphological analyzer .
  - LDB software's query facilities, which allow fast access to POS (CAT in our case) field in definitions.

- Take the greatest advantage of the information contained in the dictionary.

Let's slightly describe our definite approach. We have implemented a kind of merge between the two directions described within the third option, namely, the LDB is consulted on-line, but there is a separate lexicon where both entries not present in the dictionary and extremely irregular forms are included. The use of this separate lexicon allows to override the information provided by LDB (see [Sanfilippo 90a, 90b] to get more information about the Seg-Word analyzer).

After the basic process of categorization, which returns a list of possible POS associated to each form of the input, an additional step of Morphological Optimization is carried out. This step has been introduced to minimize the effects of ambiguities in categories, taking advantage of the particular characteristics of dictionary definitions. It uses the list of <words list-of-POS> as input and performs a detection of some patterns according to the kind of definitions being processed.Thus, these patterns will be categorized in such a way that the task of the parser becomes simpler. In addition, coordination of categories is handled in advance to reduce ambiguities in the process of parsing. Punctuation symbols which do not provide the parser with any information are eliminated.

Some examples of the process are shown next. For each one, the following information is shown :

1) the dictionary entry (with its translation to English).

2) the sense definition as appearing in the Vox dictionary.
3) the input to the analyzer (with the category of the entry).
4) output of the morphological analyzer.
5) output of the optimization.
6) some comments about the latter process.

Example 1:

1. algodón (cotton).

2. algodón
acepción:1 ** m. ** Substancia fibrosa, blanca y suave, que recubre semilla de varias plantas malváceas.

3. (N "Substancia" "fibrosa" "," "blanca" "y" "suave" "," "que" "recubre" "la" "semilla" "de" "varias" "plantas" "malváceas" ".")

4. ((N CATEGORY) (SUBSTANCIA V N) (FIBROSA ADJ) (\, PUNT) (BLANCA ADJ N) (Y CONJ) (SUAVE ADJ) (\, PUNT) (QUE PRON) (RECUBRE V) (LA DET) (SEMILLA N) (DE P) (VARIAS ADJ N) (PLANTAS N) (MALV\áCEAS ADJ N))

5. ((N CATEGORY) (SUBSTANCIA V N) (FIBROSA_BLANCA_SUAVE ADJ) (QUE PRON) (RECUBRE V) (LA DET) (SEMILLA N) (DE P) (VARIAS ADJ N) (PLANTAS N) (MALV\áCEAS ADJ N))

6. those coordinated words having in common any category (adjectives "fibrosa" "blanca" and "suave" in this case), are joined together and parsed as an unique item, avoiding the possibility of errors in the parsing due to the existence of multiple categories (after the parsing process, they will be split apart again).

Example 2:

1. carbolíneo (coal tar).

2. carbolíneo [de carbón + l. oleum , aceite]
acepción:1 ** m. ** Substancia líquida y grasa, obtenida de la destilación del alquitrán de la hulla, us. para hacer impermeable la madera.

3. (N "Substancia" "líquida" "y" "grasa" "," "obtenida" "de" "la" "destilación" "del" "alquitrán" "de" "la" "hulla" "," "us." "para" "hacer" "impermeable" "la" "madera" ".")

4. ((N CATEGORY) (SUBSTANCIA V N) (L\íQUIDA ADJ) (Y CONJ) (GRASA V ADJ N) (\, PUNT) (OBTENIDA PAL) (DE P) (LA DET)(DESTILACI\óN N) (DEL PAL) (ALQUITR\áN N) (DE P) (LA DET) (HULLA N) (\, PUNT) (US. PAL) (PARA P) (HACER V) (IMPERMEABLE ADJ N) (LA DET) (MADERA V N))

5. ((N CATEGORY) (SUBSTANCIA V N) (L\íQUIDA_GRASA ADJ) (\, PUNT) (OBTENIDA_DE PATTERN4) (LA DET) (DESTILACI\óN N) (DEL PAL) (ALQUITR\áN N) (DE P) (LA DET) (HULLA N) (\, PUNT) (US. PAL) (PARA P) (HACER V) (IMPERMEABLE ADJ N) (LA DET) (MADERA V N))

6. common patterns in the present subset are detected and categorized

consequently ("obtenida de" is a common pattern to indicate the source something is got from).Thus, the rules in the corresponding grammar will easily recognize this sort of semantic information. A new case of coordination (adjectives "líquida" and "grasa") can also be observed.

Example 3:

1. galactosa (galactose).

2. galactosa
acepción:1 ** f. ** Especie de azúcar de leche, sacárido de la glucosa ordinaria, que se obtiene tratando la lactosa con ácido sulfúrico diluido.

3. (N "Especie" "de" "azúcar" "de" "leche" "," "sacárido" "de" "la" "glucosa" "ordinaria" "," "que" "se" "obtiene" "tratando" "la" "lactosa" "con" "ácido" "sulfúrico" "diluido" ".")

4. ((N CATEGORY) (ESPECIE N) (DE P) (AZ\úCAR N) (DE P) (LECHE N INTERJ) (\, PUNT) (SAC\áRIDO N) (DE P) (LA DET) (GLUCOSA N) (ORDINARIA ADJ) (\, PUNT) (QUE PRON) (SE PRON) (OBTIENE PAL) (TRATANDO GER) (LA DET) (LACTOSA N) (CON P) (\áCIDO ADJ N) (SULF\úRICO ADJ) (DILUIDO PARTI))

5. ((N CATEGORY) (ESPECIE ESP1) (PREPO PREPO) (AZ\úCAR N) (DE P) (LECHE_SAC\áRIDO N) (DE P) (LA DET) (GLUCOSA N) (ORDINARIA ADJ) (QUE PRON) (SE PRON) (OBTIENE PAL) (TRATANDO GER) (LA DET) (LACTOSA N) (CON P) (\áCIDO ADJ N) (SULF\úRICO ADJ) (DILUIDO PARTI))

6. patterns which indicate other than ISA relations between the entry word and the genus of the definition, are found and marked as such; in this particular example, "especie de" is interpreted as a TYPE-OF relation, so that the parser never takes "especie" as the genus of the definition (as it is categorized as a noun). Two new cases of coordination, now between nouns ("leche" and "sacárido") and adjectives ("ácido" and "sulfúrico") are also treated.

As the reader can observe, some words remain "uncategorized" (those whose POS is set to "PAL"). This happens with words not appearing in the dictionary (as abbreviations), and with too complex or irregular derivatives. As particular cases appear, we treat them consequently, either by adding the word (plus its categories) to the above mentioned lexicon, or by modifying/adding morphological rules.

3.1.3. Other restrictions.

Another restriction, closely related to this one, is about the extraction of semantic information from definitions. This process involves the application of Alshawi's parser over definitions followed by a sense disambiguation process.

Alshawi's parser needs, of course, a (almost) complete categorization of words in the definitions and a set of rules to cover all the types of definitions. Both kinds of information are presently incomplete and must grow incrementally during the process, through an interactive dialogue with the user.

The disambiguation process follows, roughly, Copestake's approach in the sense that it will be activated top-down, beginning with a selected top node. The use of a set of weighted heuristics for selecting the appropiate sense is also taken from Copestake

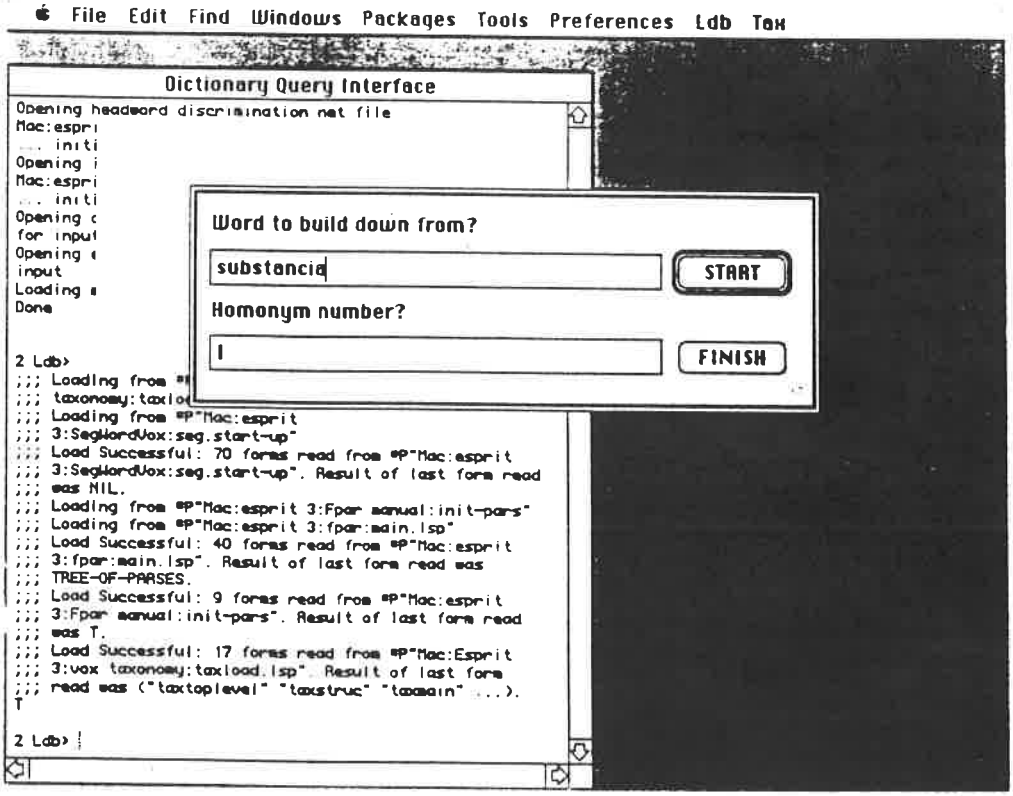
proposal but the kind of heuristics is different due to the lack of semantic codes, box codes, etc. present in Longman.

3.2. An approach to extract taxonomies from Vox dictionary.

Taking into account the restrictions expressed in the last sections, we present the methodology followed for the Vox LDB. As a brief example of the performance of the taxonomy builder, we will graphically show a session depicting the construction of a part of the "substancia" taxonomy. A more detailed description of the system capabilities will be given in a forthcoming "user manual" paper. The steps to be taken are the following:

a) Selection of possible starting points (roots) for building the taxonomy (see [Acquilex 90] to clarify the criteria followed). In our particular example we will start from the entry "substancia" .

b) The system allows the user to follow two ways starting from the top word: working at sense level or at word level. That means that we can build the taxonomic structure from a specific sense or from all the senses (or some of them). Working at word level allows a more efficient way of building large taxonomies, but states problems for the lexicographer when dealing with sense identification tasks.



Window 1.

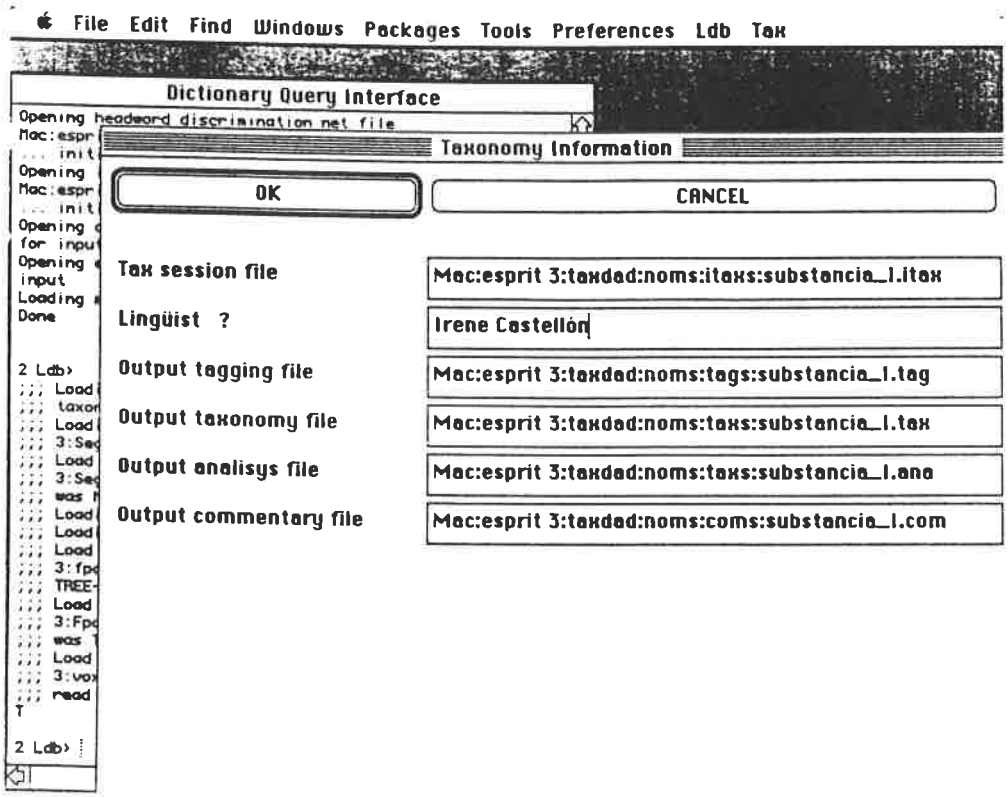
After having given the top entry "substancia" as shown in window 1, some general information about the whole taxonomy appears so that the linguist could validate default values or modify any of them (this is shown in window 2). The user must also select the analysis grammar and the heuristics to be applied.

Another possibility at this point is to merge together two or more senses or to discard some of them. At this level the merging of senses belonging to different entries is not allowed.

In this case, "substancia" has 10 different senses and the goal is to reduce this number. By amalgamation when differences between them are not relevant or by deletion when the senses are characterized with some kind of label, like GEO (geography) or USO (use). In our particular case two heuristics have been applied, one to merge and the other to discard. They are as follows:

AMALGAMATE-HEU:

"if the number of non-functional words belonging to the intersection of the definitions of two senses and not appearing in the definition of the other senses, is greater than a predefined amount and the POS is the same, then propose the amalgamation of the two senses".



Window 2.

Thus, senses 4,5,6 of "substancia" are merged together because of the presence of flexive and derivative forms of "alimento" and "nutrir".

sense:4**f.**Cosa con que otra se **alimenta** y **nutre** y sin la cual se acaba.

sense:5**f.** Parte **nutritiva** de los **alimentos**.

sense:6**f.** Jugo que se extrae de ciertas materias **alimenticias**.

SENSE_ELIM_HEU:

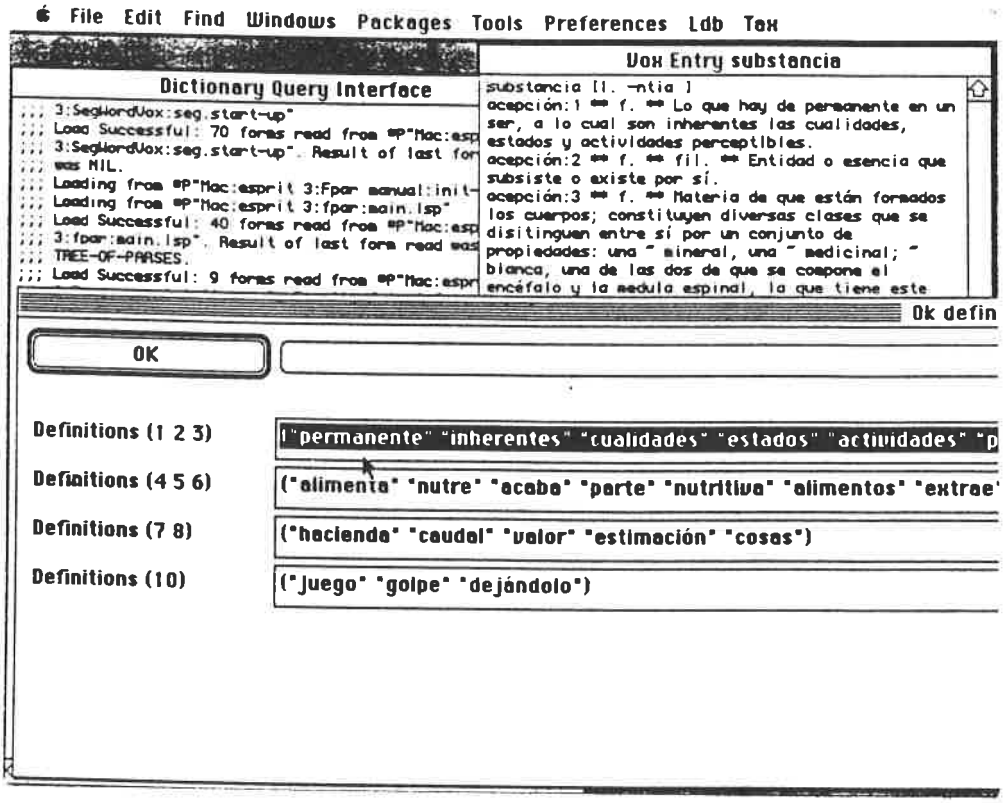
"discard senses where tags SEM, REG, USO or GEO are filled".

Hence, sense 9 of "substancia" is discarded.

sense:9 ** f. ** **fig**. ** **fam**. ** Juicio, madurez: hombre sin ~.

Following the general criteria defined above, this process is achieved taking into account a set of heuristics to be defined incrementally. Of course querying the user would be the default one (see section 4).

Significant words of each final sense definition in the top entry can as well be validated and/or modified (see window 3).



Window 3.

c) The next step consists of searching in the dictionary definitions for all the occurrences of the top word (and, perhaps, of some inflectional forms or derivatives). We will obtain, then, a list of senses.

d) For each sense two problems must be solved:

1- is the word we are searching for acting as a genus ?
2- sense disambiguation problem.

These two problems are mutually related, although usually the former can be solved separately.

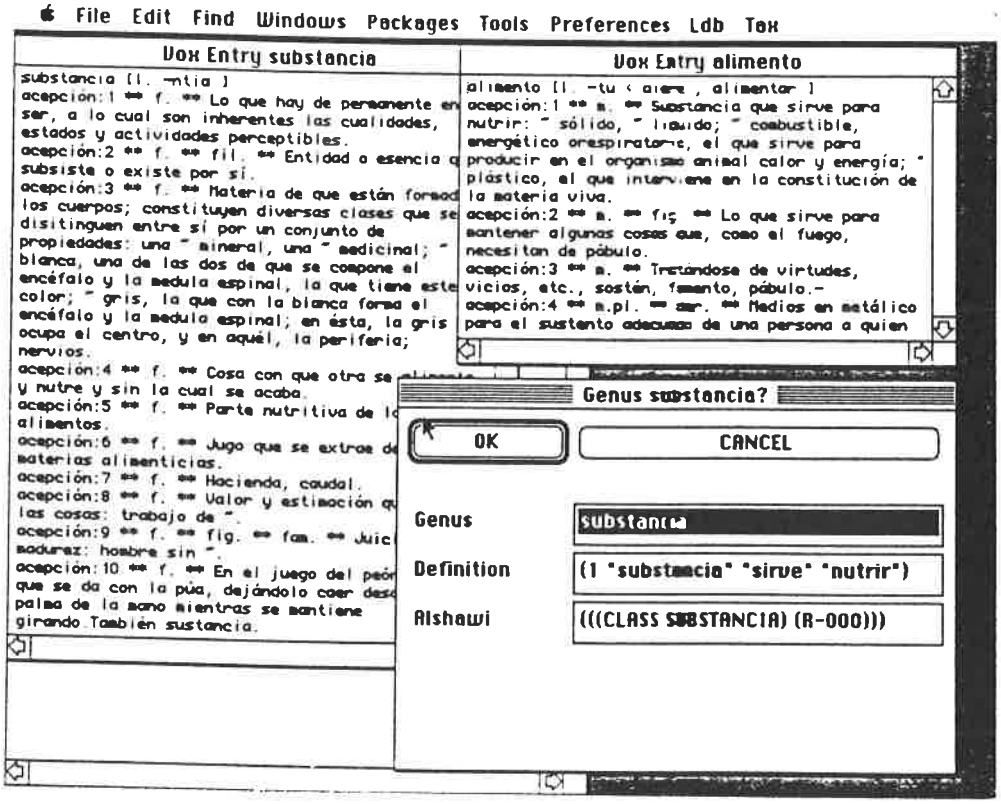
Only definitions where the word appears as genus are considered.

The result of the Alshawi parser falls on one of the following cases:

- the word has been correctly selected as genus;
- the word has been correctly discarded as genus;
- the performance of parser has not been correct due either to the lack of categorization of some words appearing in the definition or the incompleteness of the definition grammar.

The last possibility, malfunction of the parser, suggests that some kind of interactive modification, in the lexicon, in the morpho-rules or in the grammar, must be allowed.

In window 4, we can see one of these senses (*alimento (1)* "substancia que sirve para nutrir"), the result of applying the extraction of its genus "substancia", as well as the most significant words in the sense definition.At this point, if the user considers that *alimento (1)* is indeed a hyponym of substancia, he must click the OK button, correcting if necessary the most significant words and the result of the Alshawi analysis. In the other case, the user must click the CANCEL button.



Window 4.

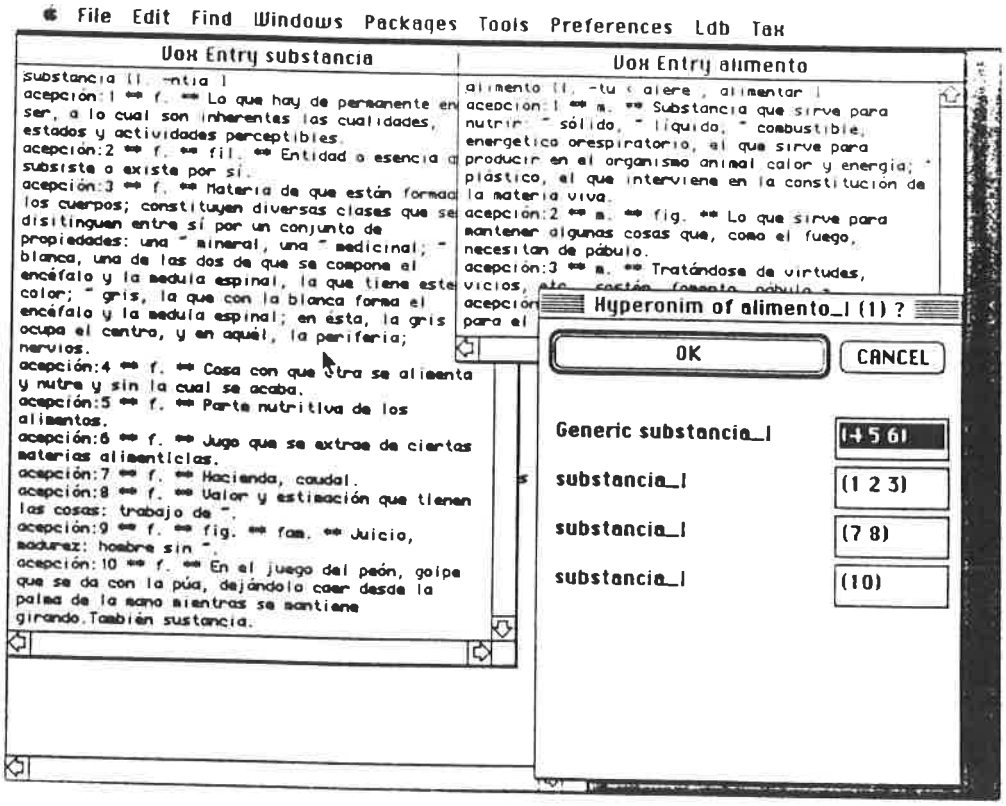
Once a genus term is selected, different kind of heuristics are applied to perform sense disambiguation. In our example, the following heuristic is applied:

SEL-PATTERN-HEU:

"prefer the greatest intersection between not functional words appearing in the definition of the current sense and words appearing in the definitions of each sense of the genus term".

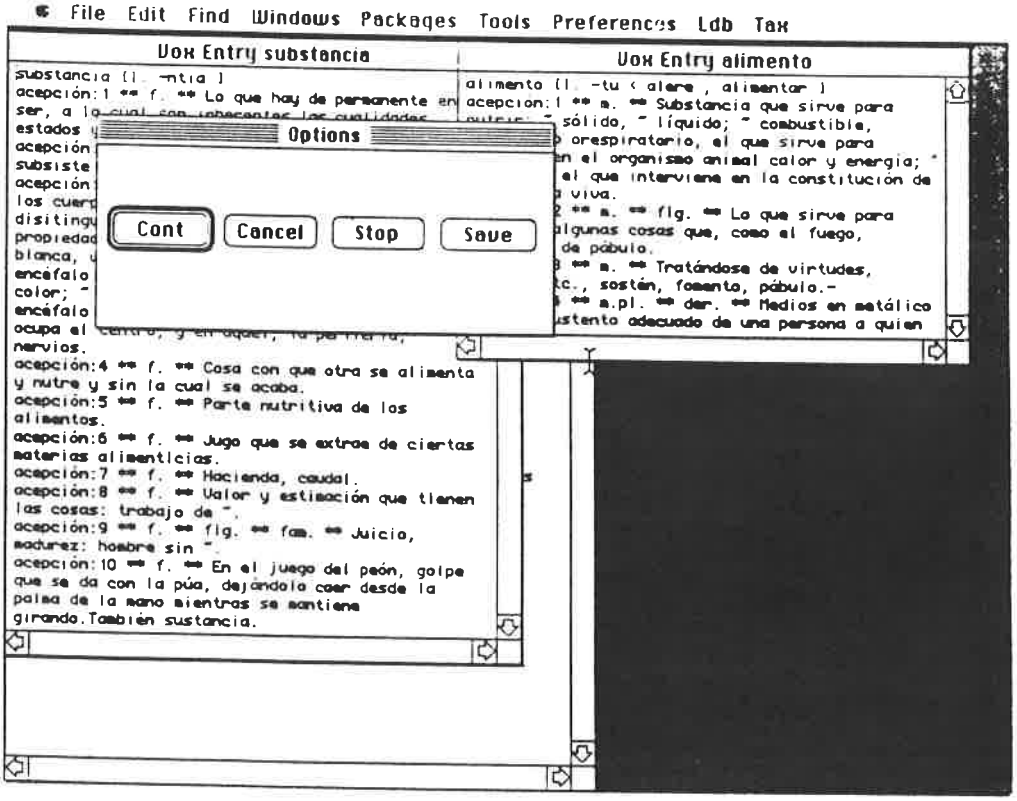
Therefore, *alimento(1)* is connected with the generic *substancia (4 5 6)* due to the occurrence of "nutrir" in both definitions (see window 5).

At this point, the user must choose among the set of possible hyperonym senses of *alimento (1)* . These appear ordered according to the resulting values of the applied heuristics. In the first place, thus, appears the sense of "substancia" which has the greatest possibility of being the autentic hyperonym of *alimento (1)* : *substancia (4 5 6).*



Window 5.

e) The process then recurs over steps c) and d). The word "alimento" is then taken as top entry. When the taxonomy process of *alimento (1)* has finished, we proceed with the next possible hyponym of "substancia".



Window 6.

In certain moments, whenever a top entry happens to appear in any definition, the user is queried about one out of four possibilities: continue to build the taxonomy (in depth with the first appearance of the top in a definition), prune this branch of the taxonomy tree (leaving the top entry in the taxonomy structure for a further completion), stop the process (previously saving the present state) or just save the present situation (returning to the same point). These options can be seen in window 6.

f) The process will stop automatically if the queue is empty. In the other case, other heuristics could decide to halt the process.

g) The process generates five different output files:

- One containing the general information (for internal use).

- Another with the taxonomy structure (also for internal use).

- Another one with the last analysis performed (in plain text):

```
substancia I (1 2 3)
    ...
substancia I (4 5 6)
    ...
    alimento I (1)
    (((CLASS SUBSTANCIA) (R-000))) T

        ahumado I (3)
        (((CLASS ALIMENTO) (R-000))) T
        ...
    ...
substancia I (7 8)
    ...
substancia I (10)
    ...
```

- The one that contains just the skeleton of the taxonomy (also in plain text):

```
substancia I (1 2 3)
    ...
substancia I (4 5 6)
    ...
    alimento I (1)
        ahumado I (3)
        ...
    ...
substancia I (7 8)
    ...
substancia I (10)
    ...
```

A more extensive example with the "bebida" taxonomy can be seen in APPENDIX A.

- The last one, with all the information collected in both the acquisition and validation modes (also in plain text). The result of the example can be seen in APPENDIX B.

4. Heuristic evaluation and management.

The definitions of sets of parametrized heuristics, the use of these sets for guiding the selection process and the existence of a mechanism for evaluating the performance and allowing the updating of such heuristics, are relevant features of our system.

For each point where the system has to make a choice, a set of heuristics must be available.

Heuristics are means of implementing criteria for taking decisions in situations where no algorithmic solution can be stated.

Basically, a heuristic is a procedure that assigns a score to each of the different options it must consider. A global score, result of those corresponding to each heuristic, is obtained, and then, a decision is taken based on these global scores.

Consider, for instance, the problem of selecting the correct sense of the hyperonym term once the hyperonym relation has been stated, as was discussed in section 3.2. We don´t have any algorithmic solution for this problem. Querying the lexicographer could be a valid alternative but, of course, we must improve it.

Although no algorithmic solution exists, there are however some, more or less loose, criteria to be applied: "select the first sense", "rank the scores of each sense in a linear way", "do not consider senses with some special property (having a geographical code, for instance)" and so forth.

We must, then build different heuristics for implementing each of the above criterium and assign a weight to each of these heuristics. Some of them are depicted in figures 5 and 6, appearing as leaves of the tree.

At running time, once the choice point is reached, the procedures attached to all of the involved heuristics are evaluated. Their results are weighted and a global result is build and taken into account for making the right choice.

The performance of a set of heuristics can, of course, be tested in order to improve their results.

In our system, this learning process can be accomplished on the validation mode: once a significant amount of cases has been analysed by the system and checked by the lexicographer, through one or more acquisition sessions, the set of heuristics involved can be modified and confronted with a test sample through validation processes. This facility allows the user to experiment with different criteria, change the parameters and weights of the heuristics and compare their performance.

Heuristics are described, in a declarative way, using a frame-oriented formalism. Frames are used to represent generic classes of heuristics or individual instances. Information is attached to frames by means of slots. Different heuristics, attached to different choice points can be classified within a hierarchical structure, linked by means of **ISA** and **INSTANCE** relations, allowing properties inheritance. A sample of this structure is shown in fig. 5.

The top of the hierarchy is the **HEURISTIC** frame, which defines the class of all heuristics.

The slots attached to **HEURISTIC** are the ones shared by all. **TYPE** is a mandatory descriptor that must be filled with the identification of the kind of choices the heuristic can accomplish. Among the possible values for this descriptor, are **SD** for **SEL-SENSE-HEURISTIC** and **ME** for **SEL-META-HEURISTIC** as fig. 5 shows.

Another mandatory descriptor is **WEIGHT**. This descriptor is used by the system (by means of meta-heuristics) for weighting the different heuristics applied for a certain choice process in order to take the final decision.

The **PRECONDITION** descriptor, when present, allows the declaration of a function to be evaluated in order to activate a heuristic.

The **INPUT** and **OUTPUT** descriptors must be used to declare the interface (both input and output arguments) of the heuristic procedure. Finally, **INTERPRETATION** must be filled with the function attached to the heuristic. These functions are, in fact, what is needed for the extraction process. The set of functions attached to different instances of **HEURISTIC** is recovered from this structure and loaded into our environment.

Below in the hierarchy, and, thus, inheriting from **HEURISTIC** all its descriptors, we can find **SELECTIVE-HEURISTIC**. This frame describes the class of heuristics that can be used to assign scores to a list of possible choices. The **OUTPUT** is, in this case, a list of scores. The **INPUT** depends on the specific kind of problem. In any case the form of the **INPUT** descriptor is a list of elements. The kind of object element must be declared by means of the **ELEMENT** descriptor. **SEL-SENSE-HEURISTIC**, for instance, has **ELEMENT** filled with the value "sense" and thus, the **INPUT** descriptor is "a list of senses". **SEL-META-HEURISTIC**, in another way, has as **INPUT** a "list of scores" and outputs an index over this list.

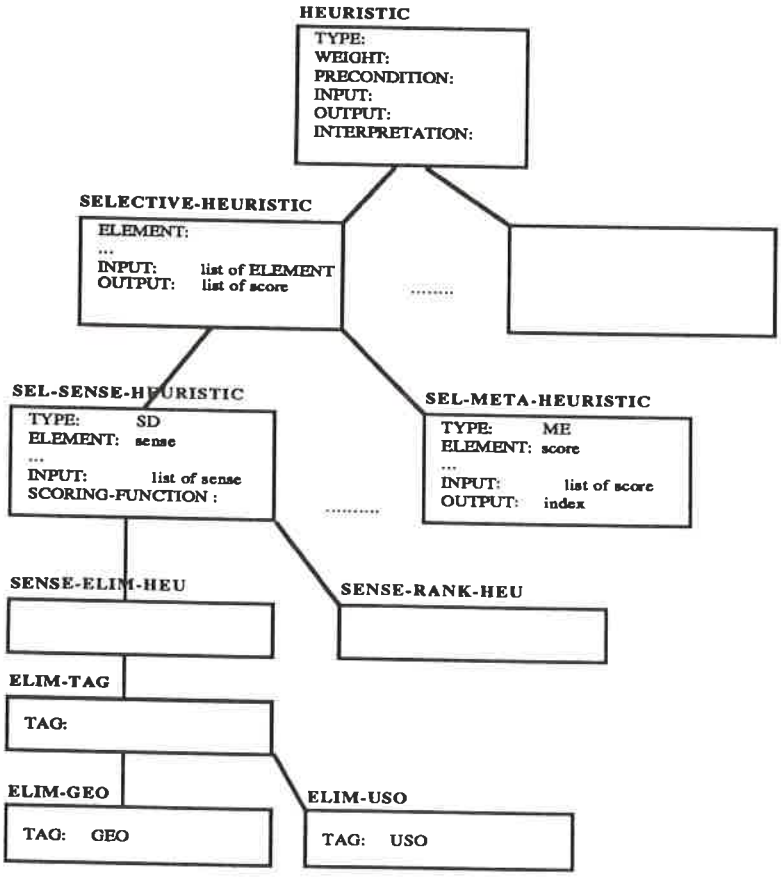


Fig. 5: Partial view of heuristics structure

In this case, the interpretation procedure is placed at **SEL-SENSE-HEURISTIC** level, although this interpretation procedure could be overridden by a more specific one,

placed lower in the tree.

Roughly speaking, this interpretation procedure consists on the application of a scoring function to the list of senses for obtaining a list of scores. The scoring function can be defined at a lower level, although a default function, at **SEL-SENSE-HEURISTIC** level could be used.

Several subclasses can be defined. For instance, **SENSE-ELIM-HEU** allows the elimination of some of the senses and **SENSE-RANK-HEU** allows the assignment of a rank of scores to the different senses.

SENSE-ELIM-HEU assigns, for instance, a score of 0 to all the deleted senses and 1 to the others. The decision of deleting or not a sense is left to a predicate to be applied to the different senses. These predicates are usually defined at a lower level. In the example of figure 5, each of the two instances, **ELIM-GEO** and **ELIM-USO**, contain a predicate that evaluates to true if the sense contains, respectively, a **GEO** or a **USO** tags.

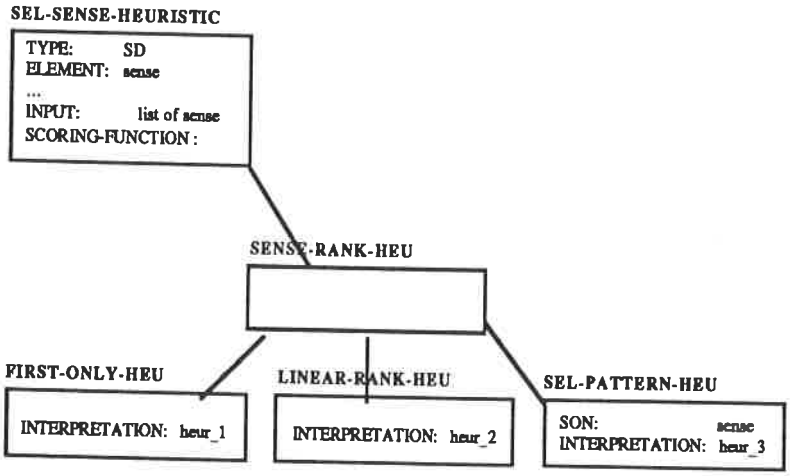


Fig. 6: Partial view of heuristics structure

In **SENSE-RANK-HEU** (see fig. 6) the scoring function ranks the scores over the full list of senses. **LINEAR-RANK-HEU**, for instance, ranks the values linearly from 1 to 0; **FIRST-ONLY-HEU** selects only the first sense, the scores are then (1,0,...,0); **SIGMOID** applies a sigmoid function, and so on...

In the **SEL-PATTERN-HEU** heuristic, the sense contained in the **SON** descriptor is compared with the list of senses (belonging to the candidates to establish a hyperonym relation). The **INTERPRETATION** procedure, in this case, compares the significative words of the corresponding definitions searching for pattern matchings between them (flexive and declarative variants are taken into account). The degree of matching is the clue of scoring.

Other similar heuristics verify the presence of the word we are classifying within the definitions of the candidate senses, look for thematic coherence and so on.

Besides this classification, individual heuristics are grouped together according to the point in the extraction process where they can be applied. We can consider, thus, heuristics for sense selection, for amalgamation of senses, for stopping the growing of a branch of the taxonomic tree and so on...

After the application of a set of heuristics to a particular problem, a decision must be taken by the system. This task is accomplished by a set of meta-heuristics. This set can consist of a single heuristic: " select the option with the highest score". More complicated criteria can be applied: For instance "if the difference between the first and

second options is less than a predefined value, ask the user for a choice; if not, select the first one"...

The assignment of scores to the different options can also be defined by means of several functions. For instance, either the average or the maximum of scores assigned by the sundry heuristics applied can be used.

As we said before, two modes of operation can be applied. In acquisition mode the user must decide whether the decision of the system is correct or not. In validation mode this test is made by comparing the results stored with those generated by the system. Anyway all the results account for every heuristic applied.

5. State of development and conclusions.

An environment for management and extraction of taxonomies from on-line dictionaries has been presented. The system was built taken into account very strict requirements, due to the need of having a prototype as soon as possible to save time. We have designed a system including some tools kindly made available to us by the other partners, combining both interactive and batch execution modes, tayloring the system to the specific characteristics of the Vox dictionary, and designing it in such a way, that the end product will be as user-friendly as possible.

An initial set of heuristics has been built mainly for dealing with sense disambiguation tasks. Different taxonomies have been constructed using this environment ("substancia" (substance), "alimento" (food), "bebida" (drink)).

The required linguistic knowledge sources (FPar grammars and Seg-Word rules) have been developed concurrently with the taxonomy building environment.

Current topics of interest include the improvement of the heuristic set and its augmentation to allow a more automatic performance of the system for extracting other taxonomies (especially verb taxonomies).The FPar grammar is being refined as well, in order to extract other types of relations and semantic properties.

Appendix A

bebida I (1 2)
bebida I (3)
  absenta I (1)
  aguardiente I (1)
    anisado I (2)
      anís I (5)
        palomita I (2)
    armañac I (1)
    balarrasa I (1)
    cachaza I (2)
    calvados I (1)
    cazalla I (1)
    coñac I (1)
      coñá I (1)
    holanda I (2)
    kirsch I (1)
    locumba I (1)
    matarratas I (2)
    ojén I (1)
    refinado I (3)
    sake I (1)
    vodka I (1)
      vodca I (1)
  ajenjo I (2)
    absenta I (1)
  almendrada I (1)
  angélica I (5)
  aperitivo I (2)
  atole I (1)
  aurora I (6)
  avenate I (1)
  batido I (5)
  bebedizo I (2)
  beber I (1)
  bebido I (2)
  bebienda I (1)
  bíter I (1)
    bítter I (1)
  brandi I (1)
  brebaje I (1)
    brebajo I (1)
    filtro II (1)
      bebedizo I (3)
  cacao I (3)
  calonche I (1)
  carraspada I (1)
    garnacha II (3)
  cedrito I (1)
  cerveza I (1)
    ale I (1)
    pelel I (1)
  chicha II (1)
    azua I (1)
  chinchón I (1)
  chocolate I (2)

suizo I (5)
clara I (7)
clarea I (1)
cordial I (4)
cuba I (5)
cubalibre I (1)
cúmel I (1)
filtro II (2)
flip I (1)
galacina I (1)
gaseosa I (1)
ginfizz I (1)
grog I (1)
guarapo I (2)
helado I (4)
polo IV (1)
queso I (3)
tutti I (1)
hipocrás I (1)
horchata I (1)
hordiate I (2)
jarabe I (1)
acetomiel I (1)
arrope I (3)
churdón I (3)
diacodión I (1)
jarope I (1)
lamedor I (3)
melito I (1)
sirope I (1)
kumis I (1)
lebení I (1)
licor I (2)
alquermes I (2)
amargo I (7)
anisete I (1)
básig I (1)
benedictino I (4)
chartreuse I (1)
crema I (7)
sabayón I (1)
curasao I (1)
curazao I (1)
estomacal I (2)
galliano I (1)
güisqui I (1)
escocés I (4)
whisky I (1)
bourbon I (1)
kummel I (1)
marrasquino I (1)
murtilla I (2)
murtina I (1)
noyó I (1)
ouzo I (1)
pipermín I (1)
poscafé I (1)
ron I (1)
trasmonta I (1)

triple I (3)
tuba II (1)
limonada I (1)
líquido I (6)
zumo I (1)
acedo I (4)
agraz I (2)
agrio I (8)
lechal I (3)
linfa I (2)
meluza I (1)
mosto I (1)
mostazo I (1)
mostillo I (1)
naranjada I (1)
pampanada I (1)
quino I (2)
goma I (4)
regaliza I (2)
rob I (1)
vino I (1)
ablución I (5)
aguapié I (1)
aguachirle I (1)
torcedura I (5)
ahumado I (4)
albariño I (1)
alicante I (3)
aloque I (2)
amontillado I (1)
amoroso I (5)
añejo I (3)
cacabelos I (1)
camedrita I (1)
cariñena I (1)
carló I (1)
cava II (4)
cebreros I (1)
cécubo I (1)
chacolí I (1)
champaña I (1)
champán II (1)
chianti I (1)
clarete I (1)
cream I (1)
dolaje I (1)
duelaje I (1)
falerno I (1)
fondillón I (2)
jerez I (1)
xerez I (1)
jerte I (1)
jumilla I (1)
lágrima I (8)
madera II (1)
málaga I (1)
malvasía I (2)
manzanilla I (12)
montilla I (1)

moriles I (1)
navalcarnero I (1)
navarra I (1)
oloroso I (2)
oporto I (1)
pajarete I (1)
pajarilla I (3)
penedés I (1)
peñafiel I (1)
pitarras I (1)
priorato II (1)
purrela I (1)
quianti I (1)
repiso I (2)
requena I (1)
reserva I (12)
ribeiro I (1)
rioja I (1)
roete I (1)
rosado I (3)
rueda II (1)
sherry I (1)
tarragona I (1)
tintilla I (1)
toro III (1)
tostadillo I (2)
transfer I (1)
trinque I (1)
utiel I (1)
valdepeñas I (1)
verdea I (1)
vinagrón I (1)
vinaza I (1)
vinazo I (1)
vinillo I (2)
yecla I (1)
zupia I (2)
mantellina I (2)
mistela I (1)
mixtela I (1)
néctar I (1)
nepente I (2)
onfacomeli I (1)
oxicrato I (1)
oxizacre I (1)
perada I (2)
poción I (1)
julepe I (1)
ponche I (1)
potaje I (4)
pulque I (1)
queimada I (1)
refresco I (2)
adiafa I (1)
agasajo I (3)
caridad I (5)
refresco I (3)
caridad I (4)
chía II (2)

granadina II (1)
granizada I (2)
granizado I (1)
granizada I (2)
mazagrán I (1)
patente I (5)
refrescamiento I (1)
sorbete I (1)
arlequín I (5)
helado I (5)
mantecado I (2)
vinagrada I (1)
regalo I (4)
sidra I (1)
soda I (2)
tila I (3)
tisana I (1)
tónico I (5)
cuasina I (1)
zarzaparrilla I (2)
zurracapote I (1)
bebida I (4)
bebida I (5)

Appendix B

Taxonomy Information File : Mac:esprit 3:taxdad:noms:itaxs:substancia_I.itax
Linguist : Irene Castellón
Time : (18 6 1991 18 48)
Top : (substancia I)
Category : NOUN
Taxonomy File (plain) : Mac:esprit 3:taxdad:noms:taxs:substancia_I.tax
Taxonomy File (intern) : Mac:esprit 3:taxdad:noms:tags:substancia_I.tag
Last Analysis File (plain) : Mac:esprit 3:taxdad:noms:taxs:substancia_I.ana
Commentary File (plain) : Mac:esprit 3:taxdad:noms:coms:substancia_I.com
1 Syntactic File : Mac:Esprit 3:grammar:genus.gram
1 Heuristic File : Mac:Esprit 3:vox taxonomy:heurist1.lsp

Correct :
Heuristic 0 has been applied 0 time(s).
Heuristic 1 has been applied 1 time(s).
Heuristic 2 has been applied 1 time(s).
Heuristic 3 has been applied 1 time(s).
Heuristic 4 has been applied 0 time(s).
Heuristic 5 has been applied 0 time(s).
Heuristic 6 has been applied 0 time(s).
Heuristic 7 has been applied 0 time(s).
Heuristic 8 has been applied 0 time(s).
Heuristic 9 has been applied 0 time(s).

Failed :
Heuristic 0 has been applied 0 time(s).
Heuristic 1 has been applied 0 time(s).
Heuristic 2 has been applied 0 time(s).
Heuristic 3 has been applied 0 time(s).
Heuristic 4 has been applied 0 time(s).
Heuristic 5 has been applied 0 time(s).
Heuristic 6 has been applied 0 time(s).
Heuristic 7 has been applied 0 time(s).
Heuristic 8 has been applied 0 time(s).
Heuristic 9 has been applied 0 time(s).

substancia I (1 2 3)
DEF (permanente inherentes cualidades estados actividades perceptibles entidad esencia subsiste existe materia están formados cuerpos constituyen diversas clases disitinguen entre)
TEM NIL
SIN NIL
MOR NIL
UNIC NIL

substancia I (4 5 6)
DEF (alimenta nutre acaba parte nutritiva alimentos extrae ciertas materias alimenticias)
TEM NIL
SIN NIL
MOR NIL
UNIC NIL

alimento I (1)
DEF (substancia sirve nutrir sólido líquido)
TEM NIL

SIN NIL
MOR NIL
UNIC NIL
HEU (#S(HEU CHECK (3 2 1) OK T))
(((CLASS SUBSTANCIA) (R-000))) T

ahumado I (3)
DEF (alimento conservado mediante utilizado darles peculiar sabor)
TEM NIL
SIN NIL
MOR NIL
UNIC NIL
HEU (#S(HEU CHECK NIL OK NIL))
(((CLASS ALIMENTO) (R-000))) T

substancia I (7 8)
DEF (hacienda caudal valor estimación cosas trabajo)
TEM NIL
SIN NIL
MOR NIL
UNIC NIL

substancia I (10)
DEF (juego golpe dejándolo)
TEM NIL
SIN NIL
MOR NIL
UNIC NIL

References

[Ageno et al. 91a]Ageno A., Cardoze S., Castellón I., Martí M. A., Rigau G., Rodríguez H., Taulé M., Verdejo M. F. "The Extraction of Semantic Information from MRDs". Universitat Politécnica de Catalunya, Barcelona.
ESPRIT BRA-3030 ACQUILEX WP NO.027

[Ageno et al. 91b]Ageno A., Cardoze S., Castellón I., Martí M. A., Rigau G., Rodríguez H., Taulé M., Verdejo M. F., forthcoming. "From LDB to LKB". Universitat Politécnica de Catalunya, Barcelona.
ESPRIT BRA-3030 ACQUILEX WP NO.028

[Ageno et al. 91c]Ageno A., Cardoze S., Castellón I., Martí M. A., Rigau G., Rodríguez H., Taulé M., Verdejo M. F., forthcoming. "A Semi-automatic Process to create LKB entries". Universitat Politécnica de Catalunya, Barcelona.
ESPRIT BRA-3030 ACQUILEX WP NO.029

[Alshawi 89]Alshawi H. "Analysing the dictionary definitions". In Boguraev B., Briscoe T. (eds) *Computational Lexicography for NLP* , chapter 7. Longman, London.

[Alsawi 90] Alshawi H. "Flexible Pattern Matching Parsing Tool (FPar). Technical Manual. Computer Laboratory, University of Cambridge.
ESPRIT BRA-3030 ACQUILEX

[Amsler 81]Amsler R. "A taxonomy for English nouns and verbs". *Proceedings of the 19th Anual Meeting of the ACL*, Stanford, California, pp 133-8.

[Acquilex 90]Acquilex. "Initial definition of the vocabulary subset". Preliminary Report. 12 Month Deliverable. Amsterdam.
ESPRIT BRA-3030 ACQUILEX

[Calzolari 91]Calzolari N. "Acquiring and Representing Semantic Information in a Lexical Knowledge Base". *Proceedings of the Workshop on Lexical Semantics*, Berkeley, USA.
ESPRIT BRA-3030 ACQUILEX WP NO.016

[Carroll 90]Carroll J. "Lexical Data Base System User Manual". Computer Laboratory, University of Cambridge.
ESPRIT BRA-3030 ACQUILEX

[Castellón et al. 90]Castellón I., Martí M. A. "Gramática del Diccionario Vox". *Proceedings of the 6th Anual Meeting of the SEPLN* . San Sebastian, Spain.

[Castellón et al. 91]Castellón I., Martí M. A., Rigau G., Rodríguez H., Verdejo M. F. "Loading the MRD into the LDB. Characteristics of Vox Dictionary". Universitat Politécnica de Catalunya, Barcelona.
ESPRIT BRA-3030 ACQUILEX WP NO.019

[Copestake 90a]Copestake A. "Building Taxonomies with disambiguated word senses". Computer Laboratory, University of Cambridge.
ESPRIT BRA-3030 ACQUILEX WP NO.008

[Copestake 90b]Copestake A. "A System for building disambiguated taxonomies: draft version". Computer Laboratory, University of Cambridge.
ESPRIT BRA-3030 ACQUILEX WP NO.012

[Ratti et al. 80]Ratti D., Saba A., Catarsi M.N., Capelli G., "Analizador morfosintáctico de textos en lengua castellana". Ed. Univ. Pisa.

[Martí 85]Martí M.A. "Un sistema d'análisi morfològica per ordinador". In *Actes del 1er Congrés de Lenguatges Naturals i Llenguatges Formals*. Barcelona.

[Martí 88]Martí M.A. "Processament informàtic del llenguatge: un sistema d'anàlisi morfològica computacional". Doctoral thesis, Universitat de Barcelona.

[Sanfilippo 90a]Sanfilippo A. "A morphological Analyser for English & Italian". Computer Laboratory, University of Cambridge.
ESPRIT BRA-3030 ACQUILEX WP NO. 004

[Sanfilippo 90b]Sanfilippo A. "Notes on Seg-Word". Computer Laboratory, University of Cambridge.
ESPRIT BRA-3030 ACQUILEX

[Vox 87]*Diccionario General Ilustrado de la Lengua Española VOX*. Ed. Biblograf S.A. Barcelona.