# SEISD: An environment for extraction of Semantic Information from on-line dictionaries

**Alicia Ageno (1) Irene Castellón(1)**
**M. A. Martí (2) German Rigau (1)**
**Francesc Ribas (1) Horacio Rodriguez (1)**
**Mariona Taulé (2) Felisa Verdejo (1)**

(1) Universitat Politècnica de Catalunya. Departament de LSI.
Pau Gargallo, 5 08028-Barcelona Spain
(2) Universitat de Barcelona. Departament de Filologia Romànica.
Gran Via de les Corts Catalanes, 585 08007-Barcelona Spain

## 1 Introduction.

Knowledge Acquisition constitutes a main problem as regards the development of real Knowledge-based systems. This problem has been dealt with in a variety of ways. One of the most promising paradigms is based on the use of already existing sources in order to extract knowledge from them semiautomatically which will then be used in Knowledge-based applications.

The Acquilex Project, within which we are working, follows this paradigm. The basic aim of Acquilex is the development of techniques and methods in order to use Machine Readable Dictionaries (MRD) * for building lexical components for Natural Language Processing Systems.

SEISD (**S**istema de **E**xtracción de **I**nformación **S**emántica de **D**iccionarios) is an environment for extracting semantic information from MRDs [Ageno et al. 91b]. The system takes as its input a Lexical Database (LDB) where all the information contained in the MRD has been stored in an structured format.

The extraction process is not fully automatic. To some extent, the choices made by the system must be both validated and confirmed by a human expert. Thus, an interactive environment must be used for performing such a task.

One of the main contribution of our system lies in the way it guides the interactive process, focusing on the choice points and providing access to the information relevant to decision taking.

System performance is controlled by a set of weighted heuristics that supplies the lack of algorithmic criteria or their vagueness in several crucial decision points.

We will now summarize the most important characteristics of our system:

• An underlying methodology for semantic extraction from lexical sources has been developped taking into account the characteristics of LDB and the intented semantic features to be extracted.

• The Environment has been conceived as a support for the Methodology.

• The Environment allows both interactive and batch modes of performance.

• Great attention has been paid to reusability. The design and implementation of the system has involved an intensive re-use of existing lexical software (written both within and outside Acquilex project). On the other hand the possibility of further use of our own pieces of software has also been taken into account.

• The system performance is controlled by a set of heuristics. The system provides us with a means of evaluating and modifying these sets in order to improve its own autonomy .

• The system has been used to extract semantic information from the Vox Spanish dictionary.

* We acknowledge the facilities received from Biblograf, S.A. for using its Vox MRD.

## 2 Methodology.

The final goal of a system like ours [Ageno et al., 91a] is to obtain a large conceptual structure where the nodes would correspond to the lexical senses in the dictionary, the information present in definitions would be encoded within the nodes and the relations would be made explicit.

The kind of relations we can set between senses are the relations that appear, in an explicit or implicit form, in the dictionary entries. The most important relation is, of course, the **ISA** one, which allows us to build a taxonomy of concepts related by the hypernym-hyponym links.

Although a brute force approach is used sometimes for limited purposes, we cannot follow this for two main reasons:

• The lack of limitations over the words that could appear in the dictionary definitions that would imply the use of a general-purpose morphological analyzer with a very large coverage.

• The need for different grammars to parse entry definitions belonging to distant semantic fields (we use different grammars for parsing entries belonging to "substance", "food" or "instrument" fields).

The conclusion was to build the whole conceptual structure from several "chunks" of conceptual nets, so that each one would correspond to a narrow domain and would be built independently. For each of these domains we have selected one or more starting words or senses (that correspond to the root of the taxonomies we intend to extract) and proceeded top-down from them.

## 3 Overview of the system.

Our system carries out four differents tasks: taxonomy construction, semantic relations extraction, heuristics

validation and knowledge integration into a LKB (Lexical Knowledge Base that will contain the conceptual structures extracted from the LDB) as shown in figure 1. The first one consists of the extraction of the taxonomy structure which underlies the dictionary definitions, starting from a top entry. The second, the extraction of the other semantic relations which appear in the definitions of the taxonomy already created. The validation of the heuristics applied in the taxonomy construction is the third task. Finally, all the information acquired is integrated into the LKB. The choosed formalism for defining LKB structures is based on a typed Feature structure (FS) system augmented with default inheritance.

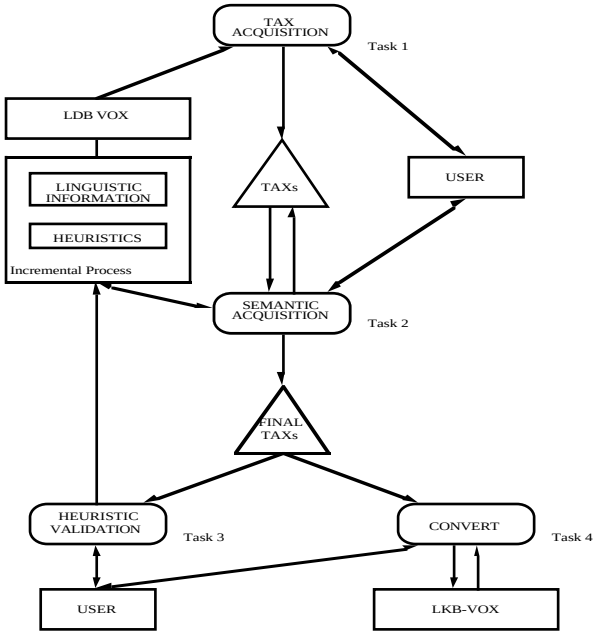


*Fig. 1: General Scheme of the System.*

### 3.1 Taxonomy Extraction.

This module is in charge of the extraction of the taxonomies which underlie the definitions of the Vox dictionary.

In our case, the problem of the extraction of the generic term is solved by means of FPar syntactic-semantic analyser [Carroll 90] with a general simplified grammar for the extraction of the generic term and specific ones for the modifiers. Given a sense, using this parser, we can detect its hyperonyms as well as other semantic relations.

The input of the analyser is a sense augmented with its morphological features.. The morphological analysis is carried out using an optimized version of Seg-Word analyzer [Sanfilippo 90].

### 3.2 Semantic Extraction.

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

The next step (semantic extraction) lies in performing a similar process to the taxonomy building, but with a different grammar and without user intervention. This batch process is called definition analysis. 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" from the definitions associated to the nodes of the taxonomy.

### 3.3 Heuristic Validation.

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, constitute relevant features of our system.

Heuristics are means of implementing criteria for taking decisions in situations where no algoritmic 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 based on these global scores is taken.

## 4 Evaluation.

The environment has been used to extract semantic information from the Vox dictionary. Vox is a monolingual Spanish dictionary containing about 90.000 entries (around 150.000 senses). We have concentrated on narrow but significative domains, including both noun ("substance", "food", "drink", "person", "place" and "instrument"), involving around 3000 senses, and verb ("movement", "ingestion" and "cooking"), involving around 300 senses, taxonomies.

An initial set of heuristics has been built mainly for dealing with sense disambiguation tasks. Different taxonomies have been constructed using this environment .

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

## References.

[Ageno et al., 91a] Ageno A., Cardoze S., Castellón I., Martí M.A., Rigau G., Rodríguez H., Taulé M., Verdejo M.F. "An environment for management and extraction of taxonomies from on-line dictionaries". UPC, Barcelona. ESPRIT BRA-3030 ACQUILEX WP NO.020

[Ageno et al. 91b]Ageno A., Cardoze S., Castellón I., Martí M. A., Ribas F., Rigau G., Rodríguez H., Taulé M., Verdejo M. F. "SEISD: User Manual". UPC, Barcelona. Research Report LSI-91-47

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

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