AN INTELLIGENT DICTIONARY HELP SYSTEM

E. Agirre, X. Arregi, X. Artola, A. Díaz de Ilarraza, F. Evrard*, K. Sarasola, A. Soroa

Informatika Fakultatea (Univ. of the Basque Country)
P. K. 649, 20080 DONOSTIA (Basque Country)
E-mail: jiparzux@si.ehu.es

(*) ENSEEIHT
2, rue Charles Canichel, 31071 Toulouse Cedex (France)

ABSTRACT.

This work discusses different issues on the construction and knowledge representation of an intelligent dictionary help system. IDHS (Intelligent Dictionary Help System) is conceived as a monolingual (explanatory) dictionary system for human use (1) (2) (3). The fact that it is intended for people instead of automatic processing distinguishes it from other systems dealing with the acquisition of semantic knowledge from conventional dictionaries. The system provides various access possibilities to the data, allowing to deduce implicit knowledge from the explicit dictionary information. IDHS deals with reasoning mechanisms analogous to those used by humans when they consult a dictionary. User level functionality of the system has been specified and a prototype has been implemented (4).

A methodology for the extraction of semantic knowledge from a conventional dictionary is briefly described. The method followed in the construction of the phrasal pattern hierarchies required by the parser (5) is based on an empirical study carried out on the structure of definition sentences. For the initial construction of the Dictionary Knowledge Base (DKB) semantic rules have been defined and attached to the syntactical patterns (6).

The representation schema proposed for the DKB (7) is basically a semantic network of frames representing word senses. After the construction of the initial DKB, several enrichment processes are performed on the DKB in order to add new facts to it; these processes are based on the exploitation of the properties of lexical-semantic relations, and also, on specially conceived deduction mechanisms. The results of the enrichment processes show the suitability of the representation schema chosen in order to deduce implicit knowledge. Erroneous deductions are mainly due to incorrect word sense disambiguation.

As an extension of IDHS, a multilingual dictionary environment is being designed on the basis of different dictionaries.

1 INTRODUCTION.

IDHS (Intelligent Dictionary Help System) is a monolingual (explanatory) dictionary system (1). Its design was conceived from the study of questions that human users would like to be answered when consulting a dictionary. The fact that it is intended for people instead of automatic processing distinguishes it from other systems dealing with the acquisition of semantic knowledge from conventional dictionaries. The system provides various access possibilities to the data, allowing to deduce implicit knowledge from the explicit dictionary information. IDHS deals with reasoning mechanisms analogous to those used by humans when they consult a dictionary.

The starting point of IDHS is a Dictionary Database (DDB) built from an ordinary French dictionary. Meaning definitions have been analysed using linguistic information from the DDB itself and interpreted to be structured as a Dictionary Knowledge Base (DKB). As a result of the parsing, different lexical-semantic relations between word senses are established by means of semantic rules (attached to the patterns); these rules are used for the initial construction of the DKB.

Once the acquisition process has been performed and the DKB built, some enrichment processes have been executed on the DKB in order to enhance its knowledge about the words in the language. Besides, the dynamic exploitation of this knowledge is made possible by means of specially conceived deduction mechanisms. Both the enrichment processes and the dynamic deduction mechanisms are based on the exploitation of the properties of the lexical semantic relations represented in the DKB (6).

The analysis of the definitions has been done after some empirical studies on the data contained in the DDB (7). The analysis mechanism is mainly based on hierarchies of phrasal patterns (5) with some extensions. The parser has been implemented, and integrated with the DDB so that the definitions are directly obtained from the DDB and the different parses result of the analysis are recorded in it. Obviously, the DDB itself has played the role of lexicon for the parser. The methodology used in the process of construction of the hierarchies is briefly explained.

In the following section an overview of IDHS is given. Section 3 presents the process of construction of the DKB. The knowledge representation model and the enrichment mechanisms are fully described in sections 4 and 5. Section 6 describes some inferential aspects of the system. In section 7, some figures about the size and contents of the prototype built are shown. Finally, in section 8, some perspectives and derived works undertaken to deal with multilingual dictionary help environments are outlined. Section 9 is devoted to present some conclusions.

2 THE IDHS DICTIONARY SYSTEM.

IDHS is a dictionary help system intended to assist a human user in language comprehension or production tasks. The architecture of IDHS includes four modules:

- The Dictionary Knowledge Base, represents by means of frame structures the knowledge extracted from the Dictionary; it has been organised in different submodules. It will be explained in more detail in section 4.
- The Inference Module. It facilitates the inferencing capabilities of the system. The basic functionality is part of this module. In section 5 more precise explanations are given.
- The Communication Module which, on the one hand, interprets the questions posed by the user, and translates them to the internal representation and, on the other hand, translates the answer of the system into a comprehensible text.
- The Interface Module which permits a friendly communication with the user.

The first two modules and a simple schema of the communication module have been specified and a prototype implemented (4). The last module is not the focus of the work here presented.

Figure 1 shows the general architecture presented.

The system provides a set of functions that have been inspired by the different reasoning processes a human user performs when consulting a conventional dictionary. Definition queries, search of alternative definitions, differences, relations and analogies between concepts, thesaurus-like word search, verification of concept properties and interconceptual relationships, etc. (8) (9) are some of the functions implemented.

For instance, a definition request, DDEF, takes as input a concept, an explanatory-level, a dictionary and a language, giving as output a definition. But it has different levels of explanation: textual (the result is just the text associated to that definition), local (the answer gives the network-like representation of the textual definition), and inherited (it produces the network-like

representation of the textual definition plus other relations deduced from the concept hierarchy). The following examples are definition queries for the meaning of *wasp* in the LPPL French dictionary, but the requested explanatory levels are different: *textual* in the first example, *local* in the second one, and *inherited* in the third one.

```
User.- DDEF (|guêpe I 1|, textual, LPPL, French, ?D)
    The user asks for the definition of wasp in French
    with "textual" as explanatory-level
System.- D= 'insecte hyménoptère à aiguillon'

U.- DDEF (|guêpe I 1|, local, LPPL, French, ?D)
    Definition of wasp in French with "local" as explanatory-level
S.- D= (and  (|guêpe I 1|  HYPERONYME       |insecte I 1|)
             (|guêpe I 1|  CARACTERISTIQUE  |hyménoptère I 1|)
             (|guêpe I 1|  POSSESSION       |aiguillon I 1|))
    Wasp is an hymenopterous insect with sting.

U.- DDEF (|guêpe I 1|, inherited, LPPL, French, ?D)
   Definition of wasp in French with "inherited" as explanatory-level.
S.-D=  (and  (|guêpe I 1|  HYPERONYME       |insecte I 1|)
             (|guêpe I 1|  CARACTERISTIQUE  |hyménoptère I 1|)
             (|guêpe I 1|  CARACTERISTIQUE  |articuler I 1#m|)
             (|guêpe I 1|  POSSESSION       |aiguillon I 1|)
             (|guêpe I 1|  POSSESSION       |patte I 1#n|)
             (|guêpe I 1|  HYPONYME         |frelon I 1|)
             (|guêpe I 1|  POSSESSEUR       |guêpier I 1|))
   Wasp is an articulated hymenopterous insect with sting and legs, a  bumblebee  is  a
   wasp, and a wasp's nest has wasps.
```

The next example will show the effects of the thesaurus-like search of concepts (RTHS). This function takes as input an expression of constraints, a dictionary, and a language, and returns the list of concepts that meet the constraints stated. Examples follow:

```
U.- RTHS((  and  (?X HYPERONYME    |instrument I 1|)
                 (?X OBJECTIF |mesurer I 1|))
            LPPL, French, ?X, ?LC)
   The user asks for nouns in French that are tools used for measurement
S.- LC=(|baromètre I 1|, |dynanomètre I 1|, |telemètre I 1| )

U.- RTHS((  and  (?X HYPERONYME |consumer I 1|)
                 (?X AGENT |feu I 1|)),
            LPPL, French, ?X, ?LC)
   The user asks for verbs in French for to consume with agent fire
S.- LC=(|brûler I 1|, |calciner I 1| )
   to burn, to blacken.
```

In summary, IDHS can be seen as a repository of dictionary knowledge apt to be accessed and exploited in several ways. The system has been implemented using KEE knowledge engineering environment.

All the knowledge represented in IDHS has been acquired from a conventional dictionary by means of parsing dictionary definitions using NLP techniques. Two different steps were distinguished when building the DKB. First the extraction of the information from the dictionary and its recording into a relational database: the Dictionary Database (DDB). This DDB was the starting point in order to create, in step 2 (see figure 2), the object oriented Dictionary Knowledge Base, that is, in fact, the support of our deduction system.

Focusing on the step 2 (construction of the DKB from the DDB) two phases are distinguished. Firstly, information contained in the DDB is used to produce an initial DKB. General information about the entries obtained from the DDB (POS, usage, examples, etc.) is conventionally

represented —attribute-value pairs in the frame structure— while the semantic component of the dictionary, i.e. the definition sentences, has been analysed and represented as an interrelated set of concepts. In this stage the relations established between concepts could still be, in some cases, of lexical-syntactic nature. In a second phase, the semantic knowledge acquisition process is completed using for that the relations established in the initial DKB. The purpose of this phase is to perform lexical and syntactical disambiguation, showing that semantic knowledge about hierarchical relations between concepts can be determinant for this.

3 BUILDING THE DICTIONARY KNOWLEDGE BASE.

The starting point of this system is a small monolingual French dictionary (*Le Plus Petit Larousse*, Paris: Librairie Larousse, 1980). This dictionary consists of nearly 23,000 senses related to almost 16,000 entries. Each entry contains the following components: part of speech (POS), meaning definition or cross-references to synonyms, marks of discourse domain usage, examples (14% of entries), etc. Among the definitions, 74% have four or less than four words. The average number of words per definition is 3.27.

The dictionary was recorded in a relational database: the Dictionary Database (DDB). This DDB is the basis of every empirical study that has been developed in order to design the final representation for the intelligent exploitation of the dictionary. The information attached in the DDB to each word occurrence in meaning descriptions was completed following a mainly automatic tagging process. Every definition word occurrence was attached to its canonical form (homograph and sense numbers included when possible). Figure 3 shows two different entries and the information associated in the database to their definition words.

The definition sentences, that is the semantic component of the dictionary, have been analyzed in the process of transformation of the data contained in the DDB to produce the DKB. The analysis mechanism used is based on hierarchies of phrasal analysis patterns (5). This mechanism seems to be especially adequate to derive and make use of partial analysis of dictionary definitions. Nevertheless, our implementation includes some modifications due mainly to its integration in the environment of the DDB.

The characterization of the different lexical-semantic relations between senses is established by means of semantic rules attached to the phrasal patterns. With regard to the construction of these semantic rules, we distinguish three types of treatment:

a) Treatment associated to definitions that follow a classic schema. The links between the *definiendum* and the *genus* are of type *subclass* and properties described by the *differentia* are expressed by means of attributes.

b) Treatment associated to synonymic definitions. In this case, an attribute representing the synonymic relation is used.

c) Treatment associated to definitions with a specific formula (specific relators). Different kinds of attributes are defined in order to represent the information conveyed by the formula.

The lexical-semantic relations between different concepts extracted from the analysis of the source dictionary are grouped into two classes:

Paradigmatic relations:

a) Synonymy and Antonymy.

b) Taxonomic relations: Hypernymy / Hyponymy (obtained from definitions of type "genus et differentia"), and Taxonymy expressed by means of specific relators such as *SORTE-DE* or *ESPECE-DE*.

c) Meronymy.

d) Others: Gradation (for adjectives and verbs), Equivalence (adjectives with past participle), Factitive and Reflexive (for verbs), Lack and Reference (to the previous sense).

Syntagmatic relations (those that relate concepts belonging to different POS's):

a) Derivation.

b) Relations between concepts without any morphological relation: case relation.

c) Others: Attributive (for verbs), Lack and Conformity.

4 REPRESENTATION OF THE DICTIONARY KNOWLEDGE: THE DKB.

The knowledge representation schema chosen for the DKB of IDHS is composed of three elements, each of them structured as a different knowledge base:

- KB-THESAURUS is the representation of the dictionary as a semantic network of frames, where each frame represents a *one-word concept* (word sense) or a *phrasal concept*. Phrasal concepts represent phrase structures associated to the occurrence of concepts in meaning definitions. Frames —or units— are interrelated by slots representing lexical-semantic relations such as synonymy, taxonomic relations (hypernymy, hyponymy, and taxonymy itself), meronymic relations (part-of, element-of, set-of, member-of), specific relations realised by means of meta-linguistic relators, casuals, etc. Other slots contain phrasal, meta-linguistic, and general information.
- KB-DICTIONARY allows access from the dictionary word level to the corresponding concept level in the DKB. Units in this knowledge base represent the entries (words) of the dictionary and are directly linked to their corresponding senses in KB-THESAURUS.
- KB-STRUCTURES contains meta-knowledge about concepts and relations in KB-DICTIONARY and KB-THESAURUS: all the different structures in the DKB are defined here specifying the corresponding slots and describing the slots by means of facets that specify their value ranges, inheritance modes, etc. Units in KB-THESAURUS and KB-DICTIONARY are subclasses or instances of classes defined in KB-STRUCTURES.

Figure 4 gives a partial view of the three knowledge bases that form the DKB with their correspondent units and their inter/intra relationships.

In the KB-THESAURUS, some of the links representing lexical-semantic relations are created when building the initial version of the knowledge base, while others are deduced later by means of specially conceived deduction mechanisms.

When a dictionary entry like ***spatule I 1:*** *sorte de cuiller plate* (*spatula: a kind of flat spoon*) is treated, new concept units are created in KB-THESAURUS (and subsidiarily in KB-DICTIONARY) and linked to others previously included in it. Due to the effect of these links new values for some properties are propagated through the resulting taxonomy.

In the example, although it is not explicit in the definition, *spatule* is "a kind of" *ustensile* and so it will inherit some of its characteristics (depending upon the inheritance role of each attribute). Fig. 4 also shows the types of concepts used: *spatule I 1* and *cuiller I 1* are noun definitions and considered subclasses of ENTITIES while *plat I 1* (an adjective) is a subclass of QUALITIES. The phrasal concept unit representing the noun phrase *cuiller plate* is treated as a hyponym of its nuclear concept (*cuiller I 1*).

4.1 KB-STRUCTURES: the meta-knowledge.

This knowledge base reflects the hierarchical organization of the knowledge included in the DKB.

We will focus on the LKB-STRUCTURES class which defines the data types used in KB-DICTIONARY and KB-THESAURUS, and that organizes the units belonging to these knowledge bases into a taxonomy.

Slots defined in KB-STRUCTURES have associated aspects such as the value class, the inheritance role determining how values in children's slots are calculated, and so on. Each lexical-semantic relation —represented by an attribute or slot— has its own inheritance role. For instance, the inheritance role of the CARACTERISTIQUE relation states that every concept inherits the union of the values of the hypernyms for that relation, while the role defined for the SYNONYMES relation inhibits value inheritance from a concept to its hyponyms.

The subclasses defined under LKB-STRUCTURES are the following:

- ENTRIES, that groups dictionary entries belonging to KB-DICTIONARY;
- DEFINITIONS, that groups word senses classified according to their POS;
- REFERENCES, concepts created in KB-THESAURUS due to their occurrence in definitions of other concepts ("definitionless");
- CONCEPTS, that groups, under a conceptual point of view, word senses and other conceptual units of KB-THESAURUS.

The classification of conceptual units under this last class is as follows:

- **TYPE-CONCEPTS** correspond to Quillian's "type nodes" (10); this class is, in fact, like a superclass under which every concept of KB-THESAURUS is placed. It is further subdivided into the classes ENTITIES, ACTIONS/EVENTS, QUALITIES and STATES, that classify different types of concepts.
- **PHRASAL-CONCEPTS** is a class that includes concepts similar to Quillian's "tokens" —occurrences of type concepts in the definition sentences—. Phrasal concepts are the representation of phrase structures that are composed by several concepts with semantic content. A phrasal concept is always built as a subclass of the class that represents its head (the noun of a noun phrase, the verb of a verb phrase, and so on), and integrated in the conceptual taxonomy. Phrasal concepts are classified into NOMINALS, VERBALS, ADJECTIVALS, and ADVERBIALS.

  For instance, |plante I 1#3| is a phrasal concept (see fig. 5), subclass of the type concept |plante I 1|, and represents the noun phrase *"une plante d'ornement" (an ornamental plant*).
- Finally, the concepts that, after the analysis phase, are not yet completely disambiguated (lexical ambiguity), are placed under the class **AMBIGUOUS-CONCEPTS**, which is further subdivided into the subclasses HOMOGRAPHE (e.g. |faculté ? ?|), SENSE (|panser I ?|), and COMPLEX (|donner I 5/6|), in order to distinguish them according to the level of ambiguity they present.

The links between units in KB-THESAURUS and KB-DICTIONARY are implemented by means of slots tagged with the name of the link they represent. These slots are defined in the different classes of KB-STRUCTURES.

The representation model used in the system is made up of two levels:

- ***Definitory level***, where the surface representation of the definition of each sense is made. Morphosyntactic features like verb mode, time, determination, etc. are represented by means of facets attached to the attributes. The definitory level is implemented using *representational attributes*. Examples of this kind of attributes are: DEF-SORTED, DEF-QUI, CARACTERISTIQUE and AVEC.
- ***Relational level***, that reflects the relational view of the lexicon. It supports the deductive behaviour of the system and is made up by means of *relational attributes*, that may eventually contain deduced knowledge. These attributes, defined in the class TYPE-

CONCEPTS, are the implementation of the interconceptual relations: ANTONYMES, AGENT, CARACTERISTIQUE, SORTE-DE, CE-QUI, etc.

4.2 KB-DICTIONARY: from words to concepts.

This knowledge base contains the links between each dictionary entry and its senses (see link 4 in fig. 4).

4.3 KB-THESAURUS: the concept network.

KB-THESAURUS stores the concept network that is implemented as a network of frames. Each node in the net is a frame that represents a conceptual unit: one-word concepts and phrasal concepts.

The arcs interconnect the concepts and represent lexical-semantic relations; they are implemented by means of frame slots containing pointers to other concepts. Hypernym and hyponym relations have been made explicit, making up a *concept taxonomy*. These taxonomic relations have been implemented using the environment hierarchical relationship, in order to get inheritance automatically.

Let us show an example. The representation of the following definition:

géranium I 1: *une plante d'ornement*

requires the creation of two new conceptual units in THESAURUS: the one which corresponds to the definiendum and the phrasal concept which represents the noun phrase of the definition. Moreover, the units which represent *plante* and *ornement* are to be created also (if they have not been previously created because their occurrence in another definition).

Let us suppose that three new units are created: |géranium I 1|, |plante I 1#3| and |ornement I 1|. Attributes in the units may contain facets (attributes for the attributes) used in the definitory level to record aspects like determination, genre and so on, but also to establish the relations between definitory attributes with their corresponding relational, or to specify the certainty that the value in a representational attribute has to be "promoted" to a corresponding relational (see below the case of the slot DE in |plante I 1#3|).

Following is given the composition of the frames of these three units at the definitory level of representation (slots are in small capitals whereas facet identifiers are in italics):

|géranium I 1|
MEMBER.OF: NOMS
GROUPE-CATEGORIEL: NOM
CLASSE-ATTRIBUT: INFO-GENERALE
TEXTE-DEFINITION: "une plante d'ornement"
CLASSE-ATTRIBUT: INFO-GENERALE
DEF-CLASSIQUE: |plante I 1#3|
CLASSE-ATTRIBUT: DEFINITOIRES
DETERMINATION: UN
GENRE: F
RELATIONNELS-CORRESPONDANTS: DEFINI-PAR

|plante I 1#3|
SUBCLASS.OF: |plante I 1|
MEMBER.OF: NOMINALES
TEXTE: "plante d'ornement"
CLASSE-ATTRIBUT: INFO-GENERALE
DE: |ornement I 1|
CLASSE-ATTRIBUT: SYNTAGMATIQUES
RELATIONNELS-CORRESPONDANTS: ORIGINE, POSSESSEUR, MATIERE, OBJECTIF
OBJECTIF: 0.9

|ornement I 1|
MEMBER.OF: REFERENCES

Before showing the representation of these units at the relational level, it has to be said that after the initial DKB has been built some deductive procedures have been executed: e.g. deduction of

inverse relationships, taxonomy formation, etc. It is to say that in fig. 5, where the relational view is presented, the relations deduced by these procedures are also represented.

The conceptual units in THESAURUS are placed in two layers (see Fig. 5), recalling the two planes of Quillian (10). The upper layer corresponds to type concepts, whereas in the lower, phrasal concepts are placed. Every phrasal concept is placed in the taxonomy directly depending from its nuclear concept, as a hyponym of it.

It is interesting to notice in the figure the relation of *conceptual equivalence* established between |géranium I 1| and |plante I 1#3| (link labelled (3)). These units represent, in fact, the same concept because |plante I 1#3|, standing for *"une plante d'ornement"*, is the definition of |géranium I 1|.

The frame of |géranium I 1| at the relational level of representation takes the following aspect, once the relational attributes have been (partially) completed:

|géranium I 1|
SUBCLASS.OF: ENTITES, |plante I 1|
MEMBER.OF: NOMS
GROUPE-CATEGORIEL: NOM
CLASSE-ATTRIBUT: INFO-GENERALE
TEXTE-DEFINITION: "une plante d'ornement"
CLASSE-ATTRIBUT: INFO-GENERALE
DEF-CLASSIQUE: |plante I 1#3|
CLASSE-ATTRIBUT: DEFINITOIRES
DETERMINATION: UN
GENRE: F
RELATIONNELS-CORRESPONDANTS: DEFINI-PAR
DEFINI-PAR: |plante I 1#3|
CLASSE-ATTRIBUT: RELATIONNELS
INVERSES-CORRESPONDANTS: DEFINITION-DE
OBJECTIF: |ornement I 1|
CLASSE-ATTRIBUT: RELATIONNELS
INVERSES-CORRESPONDANTS: OBJECTIF+INV

Let us show now another example. It is the case of two definitions stated by means of two different stereotyped formulae belonging to the lexicographic meta-language. Many verbs in the LPPL are defined by means of a formula beginning with *"rendre"* and many nouns with one beginning with *"qui"*. The definitions selected for this example correspond to the entries ***publier I 1*** and ***ajusteur I 1***, which are represented at the definitory level using the meta-languge attributes DEF-RENDRE and DEF-QUI respectively:

publier I 1: *rendre public* *(publish: to make public)*
ajusteur I 1: *qui ajuste des pièces de métal* *(metalworker:who adjusts pieces of metal)*

The frame corresponding to |publier I 1| is the following:

|publier I 1|
MEMBER.OF: VERBES
GROUPE-CATEGORIEL: VERBE
CLASSE-ATTRIBUT: INFO-GENERALE
TEXTE-DEFINITION: "rendre public"
CLASSE-ATTRIBUT: INFO-GENERALE
DEF-RENDRE: |public I 1|
CLASSE-ATTRIBUT: DEFINITOIRES
RELATIONNELS-CORRESPONDANTS: RENDRE

where it can be seen that no phrasal concept is involved because the link (DEF-RENDRE) is established directly between |publier I 1| and |public I 1|. However, in the case of the definition of *ajusteur I 1*, two phrasal concepts are created: the attribute DEF-QUI points to the phrasal concept |ajuster I 1#1|, representing *"ajuster des pièces de métal"*, and this phrasal concept, in turn, has a syntagmatic attribute (OBJET) pointing to a nominal that represents *"pièce de métal"*. Let us show the frames involved in this last case:

|ajusteur I 1|
MEMBER.OF: NOMS
GROUPE-CATEGORIEL: NOM
CLASSE-ATTRIBUT: INFO-GENERALE

TEXTE-DEFINITION: "qui ajuste des pièces de métal"
CLASSE-ATTRIBUT: INFO-GENERALE
DEF-QUI: |ajuster I 1#1|
CLASSE-ATTRIBUT: DEFINITOIRES
MODE: IND
ASPECT: NT
TEMPS: PRES
PERSONNE: 3
RELATIONNELS-CORRESPONDANTS: QUI

|ajuster I 1#1|
SUBCLASS.OF: |ajuster I 1|
MEMBER.OF: VERBALES
TEXTE: "ajuster des pièces de métal"
CLASSE-ATTRIBUT: INFO-GENERALE
OBJET: |pièce I 1#2|
CLASSE-ATTRIBUT: SYNTAGMATIQUES
DETERMINATION: UN
NOMBRE: PL
RELATIONNELS-CORRESPONDANTS: THEME

|pièce I 1#2|
SUBCLASS.OF: |pièce I 1|
MEMBER.OF: NOMINALES
TEXTE: "pièce de métal"
CLASSE-ATTRIBUT: INFO-GENERALE
DE: |métal I 1|
CLASSE-ATTRIBUT: SYNTAGMATIQUES
RELATIONNELS-CORRESPONDANTS: ORIGINE, POSSESSEUR, MATIERE, OBJECTIF
MATIERE: 0.9

Frequently, phrasal concepts represent "unlabelled" concepts, i.e., they indeed represent concepts that do not have a significant in the language. For instance, there is not, at least in French, a verbal concept meaning *'ajuster des pièces de métal'* nor a noun meaning *'pièce de métal'*. However, this is not the case of the phrasal concepts that are linked to type concepts by means of the relation DEFINI-PAR/DEFINITION-DE, because there, the phrasal concept is, in fact, another representation of the concept being defined (see above the example of the definition of *géranium I 1*). In the representation model proposed in this work, phrasal concepts denote concepts that are typically expressed in a periphrastic way and that do not have necessarily any corresponding entry in the dictionary[1].

Another interesting point related to the creation of these phrasal concepts is the maintenance of direct links between a concept and all the occurrences of this concept in the definition sentences of other concepts. It gives, in fact, a virtual set of usage examples that may be useful for different functions of the final system.

5 ENRICHMENT PROCESSES PERFORMED ON THE DKB.

In this section the enrichment processes accomplished on the DKB are explained. Two phases are distinguished: (a) the enrichment obtained during the construction of the initial DKB, and (b), where different tasks concerning mainly the exploitation of the properties of synonymy and taxonymy have been performed.

5.1 Enrichment obtained during the construction of the initial DKB.

KB-THESAURUS itself, represented —as a network— at the relational level, can be considered an enrichment of the definitory level because, while the DKB was built, the following processes have been performed:

- Values coming from the definitory level have been promoted to the relational level.
- Values coming from the unit representing the definiens have been transferred to the corresponding definiendum unit.
- The maintenance of the relations in both directions has been automatically guaranteed.
- The concepts included in REFERENCES have been directly related to other concepts.

- The taxonomy of concepts has been made explicit, thus obtaining value inheritance.

5.2 Second phase in the enrichment of the DKB.

Several processes have been carried out in order to infer new facts to be asserted in the DKB[2]. The enrichment obtained in this phase concerns the two following aspects:

- Exploitation of the properties of synonymy (symmetric and transitive).
- Enlargement of the concept taxonomy based on synonymy.

Another aspect that has been considered to be exploited in this phase is that of disambiguation. The use of the lexical-semantic knowledge about hierarchical relations contained in the DKB can be determinant in order to reduce the level of lexical and syntactical ambiguity[3]. Heuristics based on the taxonomic and synonymic knowledge obtained previously have been considered in this phase. Some of them have been designed, implemented and evaluated in a sample of the DKB.

6 INFERENTIAL ASPECTS: DYNAMIC DEDUCTION OF KNOWLEDGE.

Dynamic acquisition of knowledge deals with the knowledge not explicitly represented in the DKB and captured by means of especially conceived mechanisms which are activated when the system is to answer a question posed by the user (8). The following aspects are considered:

- Inheritance (concept taxonomy).
- Composition of lexical relations.
- Links between concepts and relations: users are allowed to use actual concepts to denote relationships (and not only primitive relations).
- Ambiguity in the DKB: treatment of remaining uncertainty.

In the following, some aspects concerning to the second point will be discussed.

In IDHS, the relationships among the different lexical-semantic relations can be easily expressed in a declarative way. It is the way of expressing these relationships that is called the *composition of lexical relations*. From an operative point of view, this mechanism permits the dynamic exploitation —under the user's requests— of the properties of the lexical relations in a direct manner. It is, in fact, a way of acquiring implicit knowledge from the DKB.

The declarative aspect of the mechanism is based on the definition of triples: each triple expresses a relationship among different lexical-semantic relations. These triples have the form (R_1 R_2 R_3), where R_i represents a lexical relation[4]. The operative effect of these declarations is the dynamic creation of transitivity rules based on the triples stated. The general form of these rules is the following:

if X R_1 Y **and** Y R_2 Z **then** X R_3 Z

When the value(s) of the attribute R_3 are asked, a reading demon (attached to the attribute) creates the rule and fires the reasoning process with a backward-chaining strategy. The deduced facts, if any, will not be asserted in the background of the DKB, but in a temporary context.

For instance, the problem of transitivity in meronymic relations (Cruse, 86; Winston *et al.*, 87) can be easily expressed by stating the triple (PARTIE-DE PARTIE-DE PARTIE-DE) but not stating, for instance, (PARTIE-DE MEMBRE-DE PARTIE-DE), thus expressing that the transitivity in the second case is not true. Examples of other triples that have been stated in the system are:

- Combination of meronymic and non-meronymic relations:

  (PARTIE-DE LOCATIF LOCATIF)
  (LOCATIF HYPERONYME LOCATIF)
  (MEMBRE-DE HYPERONYME MEMBRE-DE)

- Combination of relations derived from the definition meta-language:
  (CARACTERISTIQUE QUI-A POSSESSION)
  (OBJECTIF CE-QUI OBJECTIF)

Explicit rules of lexical composition can be used when the general form of the triples is not valid. These rules are used following the same reasoning strategy.

Following is given the rule derived from the last triple and one instance of it. By means of this rule instance, the fact that the purpose of a *géranium* is the action of *orner* is deduced from the definitions of *géranium* and *ornement*:

if X OBJECTIF Y **and** **;;;** the objective of X is Y (entity)
Y CE-QUI Z **;;;** Y "est ce qui" Z (action)
then X OBJECTIF Z **;;;** the objective of X is Z (action)

if |géranium I 1| OBJECTIF |ornement I 1| **and**
|ornement I 1| CE-QUI |orner I 1|
then |géranium I 1| OBJECTIF |orner I 1|

7 THE PROTOTYPE OF IDHS: SIZE AND CONTENTS OF THE DKB.

The prototype obtained after the construction of the DKB contains an important subset of the source dictionary. The quality of the semantic knowledge extracted from the DDB is conditioned by the size of definitions in the dictionary. In our case definitions are short and many of them use no more than one, two or three synonyms.

KB-DICTIONNAIRE contains 2400 entries, each one representing one word. KB-THESAURUS contains 6130 conceptual units; 1738 units of these are phrasal concepts. In this KB there are 1255 ambiguous concepts. Once the initial construction phase was finished, 19691 relational arcs —interconceptual relationships— had been established. After the enrichment processes, the number of relational links have been incremented up to 21800 (10.7% more). It has been estimated that, using the mechanism of lexical composition, the number of interconceptual relations could reach an increment of between 5 and 10%[5].

Manual evaluation of a meaningful sample of 100 concept-relation-concept triples from the enriched KB-THESAURUS gave us a correctness rate of 90% (under a 95% confidence given by the size of the sample).

Concerning the deduction of semantic knowledge, two considerations arise. Firstly, the use of dubious lexical rules, such as the transitivity of synonymy, has led to some errors in the prototype. Secondly, lexical ambiguity restricts deduction, because we make ambiguous concepts stop deduction both in the enrichment process and in lexical composition. Lexical disambiguation is not a trivial issue, and is receiving much attention in recent research. Our group has developed a knowledge-based technique for lexical disambiguation of free-running text (11), which is now being applied to dictionary definitions.

8 PERSPECTIVES: A MULTILINGUAL DICTIONARY HELP SYSTEM.

Currently a multilingual environment is being designed on the basis of different dictionaries. MLDS (MultiLingual Dictionary System, an extension of IDHS) is conceived as an intelligent help system for human translators (12) (13), where two monolingual dictionaries (French and Basque) constitute the knowledge base along with a bilingual dictionary that establishes equivalence-links among concepts from the monolingual dictionaries. This allows the system to enrich its functionality as will be shown next.

As a result from our analysis of translators' needs the functions have been classified according to three main activities: source text understanding, object text generation, and search for translation

equivalents. The functions included in the monolingual dictionary help system (IDHS) give an answer to the two first activities while searching for translation equivalents would correspond to the specific functionality of the Multilingual Dictionary Help System.

There are some well known problems with lexical gaps when (a) there is no single word in the target language to express the source concept, which can be solved giving *phrasal concept equivalents*, and when (b) the source concept does not appear as an entry in the bilingual dictionaries; in this case, in order to express that the concept in the result is *more general* or *more specific* than the source concept, set operators as $\supseteq$ and $\subseteq$ can be used.

In the first two examples below there is no problem when translating the concept |*accusatif I 1*| or |*coup_de_bec I 1*| from French into Basque. In the third and fourth examples |*pattar I 1*| and |*txakolin I 1*| are not in the bilingual dictionary, so the system gives the closest concept from the monolingual dictionary and indicates whether it is more or less specific. In the last example there is no single word to say *abere* (domestic animal) in French, therefore a phrasal concept is returned.

```
T.- EQUIV ((|accusatif I 1|, , ), Basque, gram, ?LP)
S.- LP = ( (|akusatibo I 1|, , ) )

T.- EQUIV ((|coup_de_bec I 1|, , ), Basque, common, ?LP)
S.- LP = ( (|mokokada I 1|, , ) )

T.- EQUIV ((|pattar I 1|, , ), French, common, ?LP)
S.- LP =  ( (⊆,|eau-de-vie I 1|, , ) )

T.- EQUIV ((|txakolin I 1|, , ), French, common, ?LP)
S.- LP =  ( (⊇,|vin I 1|, , ) )

T.- EQUIV (|(abere I 1|, , ), French, common, ?LP)
S.- LP =  ( (|animal I 1#n|, , ) )
    where |animal I 1#n| represents "domestic animal".
```

9 CONCLUSIONS.

A methodology for the extraction of semantic knowledge from a conventional dictionary has been described. This extraction has been founded on a systematic study of dictionary definitions. A parser based on phrasal pattern hierarchies has been implemented and used in that study.

The method followed in the construction of the hierarchies needed by the parser is based on an empirical study on the structure of definition sentences. The results of its application to a real dictionary has shown that the parsing method is particularly suited to the analysis of short definition sentences, as it was the case of the source dictionary.

As a result of this process, the characterization of the different lexical-semantic relations between senses —which is the basis for the proposed DKB representation schema— has been established.

A frame-based knowledge representation model has been described. This model has been used in an Intelligent Dictionary Help System to represent the lexical knowledge acquired automatically from a conventional dictionary.

The characterisation of the different interconceptual lexical-semantic relations is the basis for the proposed model and it has been established as a result of the analysis process carried out on dictionary definitions.

Several enrichment processes have been performed on the DKB —after the initial construction— in order to add new facts to it; these processes are based on the exploitation of the properties of lexical-semantic relations. Moreover, a mechanism for acquiring —in a dynamic

way— knowledge not explicitly represented in the DKB is proposed. This mechanism is based on the composition of lexical relations.

The general objective of IDHS is to assist a human user in language comprehension or production tasks. As a particular application, IDHS has been used in the design and implementation of a computerised translation-oriented dictionary that helps human translators choosing suitable target lexical units that correspond with those that are in the source text (13). A new lexical knowledge base was constructed for Basque following the same architecture, and the IDHS functionality was enriched with the treatment of knowledge about the process of lexical translation.

REFERENCES.

(1) Artola X. (1993). "HIZTSUA: Hiztegi-sistema urgazle adimendunaren sorkuntza eta eraikuntza / Conception d'un système intelligent d'aide dictionnariale (SIAD)" Ph.D. Thesis. University of the Basque Country UPV-EHU. Donostia.

(2) Artola X., F. Evrard. (1992). Dictionnaire intelligent d'aide à la compréhension, *Actas IV Congreso Int. EURALEX´90* (Benalmádena), pp. 45-57. Barcelona: Biblograph.

(3) Agirre E., X. Arregi, X. Artola, A. Díaz de Ilarraza, K. Sarasola, A.Soroa (1996). Constructing an intelligent dictionary help system. Natural Language Engineering 2(3) pp. 229-252

(4) Agirre E., X. Arregi, X. Artola, A. Díaz de Ilarraza, F. Evrard, K. Sarasola (1994a). Intelligent Dictionary Help System. *Applications and implications of current Language for Special Purposes research.* Vol I pp. 174-183 Bergen.

(5) Alshawi, H. (1989). Analysing dictionary definitions in B. Boguraev, T. Briscoe eds., pp. 153-169, Computational Lexicography for Natural Language Processing. New York: Longman.

(6) Agirre E., X. Arregi, X. Artola, A. Díaz de Ilarraza, F. Evrard, K. Sarasola (1994b). Lexical Knowledge Representation in an Intelligent Dictionary Help System. *Proc. of COLING'94*, 544-550. Kyoto (Japan).

(7) Agirre E., X. Arregi, X. Artola, A. Díaz de Ilarraza, F. Evrard, K. Sarasola (1994c). A methodolgy for the extraction of semantic knowledge from dictionaries using phrasal patterns. Proc. of IBERAMIA'94, 263-270. Caracas (Venezuela).

(8) Arregi X., X. Artola, A. Díaz de Ilarraza, F. Evrard, K. Sarasola (1991). Aproximación funcional a DIAC: Diccionario inteligente de ayuda a la comprensión, *Proc. SEPLN*, 11, pp. 127-138.

(9) Agirre E., X. Arregi, X. Artola, A. Díaz de Ilarraza, F. Evrard, K. Sarasola (1995). IDHS, MLDS: Towards Dictionary Help Systems for Human Users. in K. Korta, J.M. Larrazabal eds.,pp167-188, Semantics and Pragmatics of Natural Language: Logical and Computational Aspects (ISBN 84-920104-3-6). ILCLI Series, no. 1, Donostia (Basque Country).

(10) Quillian M.R. (1968). Semantic Memory in M. Minsky ed., pp. 227-270, *Semantic Information Processing.* Cambridge (Mass.): MIT Press.

(11) Agirre E., Rigau G. (1996). Word Sense Disambiguation using Conceptual Density. *Proc. of COLING'96*. Copenhagen (Denmark).

(12) Arregi X. (1995). "ANHITZ: Itzulpenean laguntzeko Hiztegi-sistema eleanitza / ANHITZ: Multilingual dictionary help system for translation tasks " Ph.D. Thesis. University of the Basque Country UPV-EHU. Donostia.

(13) Agirre E., X. Arregi, X. Artola, A. Díaz de Ilarraza, H. Patel, K. Sarasola, A. Soroa (1996). A Computerised Translation-Oriented Dictionary. Proc. of NLP+IA / TAL+AI 96. Moncton (Canada)

Other references:

Agirre E., X. Arregi, X. Artola, A. Díaz de Ilarraza, K. Sarasola (1995). Lexical-semantic Information and Automatic Correction of Spelling Errors in K. Korta, J.M. Larrazabal eds.,pp157-166, Semantics and Pragmatics of Natural Language: Logical and Computational Aspects (ISBN 84-920104-3-6). . ILCLI Series, no. 1, Donostia (Basque Country).

Amsler, R.A. (1981). A. Taxonomy for English Nouns and Verbs, *Proc. 19th Annual Meeting ACL*, pp. 133-138.

Arango Gaviría, G. (1983). *Une approche pour amorcer le processus de compréhension et d'utilisation du sens des mots en langage naturel.* Thèse de 3ème cycle (Paris VI). Publications du Groupe de Recherche Claude François Picard.

Boguraev B., T. Briscoe eds. (1989). *Computational Lexicography for Natural Language Processing.* New York: Longman**.**

Byrd R.J., N. Calzolari, M.S. Chodorow, J.L. Klavans, M.S. Neff, O.A. Rizk (1987). Tools and Methods for Computational Lexicography, *Computational Linguistics* 13, 3-4, pp. 219-240.

Calzolari, N. (1984). Machine-readable dictionaries, lexical data bases and the lexical system, *Proc. COLING* (Standford Univ.), p. 460.

Calzolari N., E. Picchi (1988). Acquisition of semantic information from an on-line dictionary, *Proc. COLING* (Budapest), pp. 87-92.

Chodorow M.S., R.J. Byrd (1985). Extracting semantic hierarchies from a large on-line dictionary*, Proc. ACL*, pp. 299-304.

Chouraqui E., E. Godbert (1989). Représentation des descriptions définies dans un réseau sémantique, *Actes 7ème Congrès Reconnaissance des Formes et Intelligence Artificielle* (AFCET-INRIA, Paris), pp. 855-868.

Copestake, A. (1990). An approach to building the hierarchical element of a lexical knowledge base from a machine-readable dictionary, paper read at *First Int. Workshop on Inheritance in NLP* (Tilburg).

Cruse D.A. (1986). *Lexical Semantics.* Cambridge: Cambridge University Press.

van den Hurk I., W. Meijs. The dictionary as a corpus: analyzing LDOCE's definition-language, *Corpus Linguistics II*, pp. 99-125.

Litkowsky K.C. (1978). Models of the semantic structure of dictionaries, *American Journal of Computational Linguistics,* Mf. 81, pp. 25-74.

Markowitz J., T. Ahlswede, M. Evens (1986). Semantically significant patterns in dictionary definitions, *Proc. 24th Annual Meeting ACL* (New York), pp. 112-119.

McRoy, S. (1992). Using Multiple Knowledge Sources for Word Sense Discrimination, *Computational Linguistics*, vol. 18, num. 1.

Pazienza M.T., P. Velardi (1987). A structured representation of word-senses for semantic analysis, *Proc. 3rd European Conference ACL* (Copenhaguen), pp. 249-257.

Tsurumaru H., T. Hitaka, S. Yoshida (1986). An attempt to automatic thesaurus construction from an ordinary japanese language dictionary, *Proc. COLING* (Bonn), pp. 445-447.

Vossen P., W. Meijs, M. den Broeder (1989). Meaning and structure in dictionary definitions in B. Boguraev, T. Briscoe eds., pp. 171-192, *Computational Lexicography for Natural Language Processing*. New York: Longman.

Wilks Y., D. Fass, G. Cheng-Ming, J.E. McDonald, T. Plate, B.M. Slator (1990). Providing Machine Tractable Dictionary Tools, *Machine Translation,* no. 5, pp. 99-154.

Winston M.E., R. Chaffin, D. Herrmann (1987). A Taxonomy of Part-Whole Relations, *Cognitive Science,* no. 11, pp. 417-444.

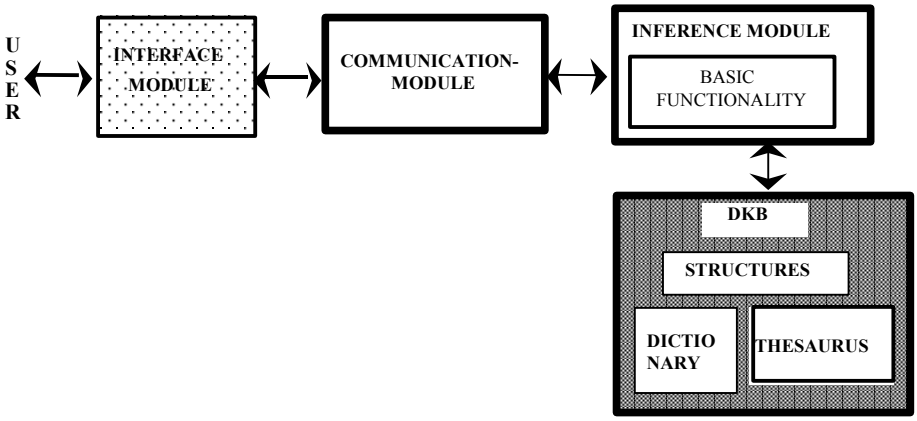

Fig. 1.- Basic Architecture of IDHS.

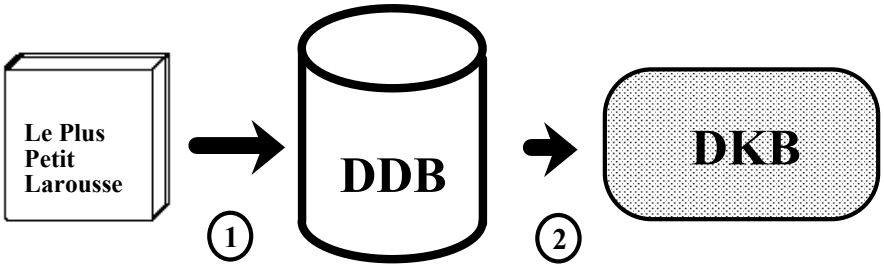

Fig. 2.- From the MRD to the DKB.

spatule I 1:	*sorte*	*de*	*cuiller*	*plate*		**(1)**
	sorte I	*de I*	*cuiller I 1*	*plat I*		**(2)**
	f.	prép.	f.	adj.		**(3)**
				F		**(4)**
	spatula: a kind of flat spoon					**(5)**
bolide I 1:	*véhicule*	*qui*	*va*	*très*	*vite*	**(1)**
	véhicule I 1	*qui I*	*aller I*	*très I 1*	*vite I 1*	**(2)**
	m.	pron. rel.	vi.	adv.	adv	**(3)**
			PI3			**(4)**
	racing car: vehicle that goes very fast					**(5)**

Fig. 3.- Two different entries in the DDB after tagging and disambiguation.
(1) definition text
(2) canonical forms
(3) POS (where ***f*** is attached to feminine nouns, ***m*** to masculin nouns, ***vi*** to intransitive verbs and **pron. rel.** to relative pronouns)
(4) ortho-morphological alterations (where ***F*** means feminine of an adjective and ***PI3*** means third person singular of present tense)
(5) English gloss

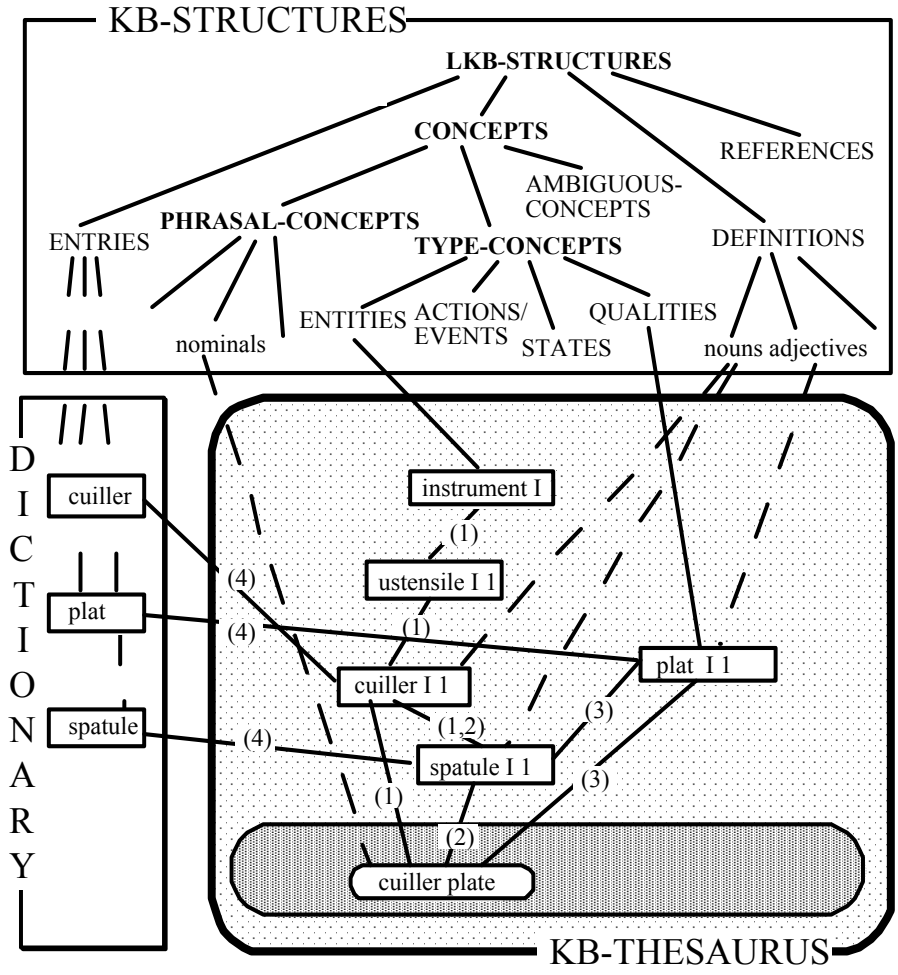

Fig. 4.- The Dictionary Knowledge Base.

—— SUBCLASS link
- - - MEMBER-OF link (instance)
(1) Taxonomic Relation: HYPERNYM/HYPONYM
(2) Specific (meta-linguistic) relation: SORTE-DE /SORTE-DE+INV (KIND-OF/KIND-OF+INV)
(3) CARACTERISTIQUE /CARACTERISTIQUE+INV (PROPERTY/PROPERTY+INV) relation
(4) MOTS-ENTREE /SENS (ENTRY-WORD / WORD-SENSE) relation
(English gloss: cuiller=spoon, cuiller plate=flat spoon, plat=flat, spatule=spatula, ustensile=instrument)

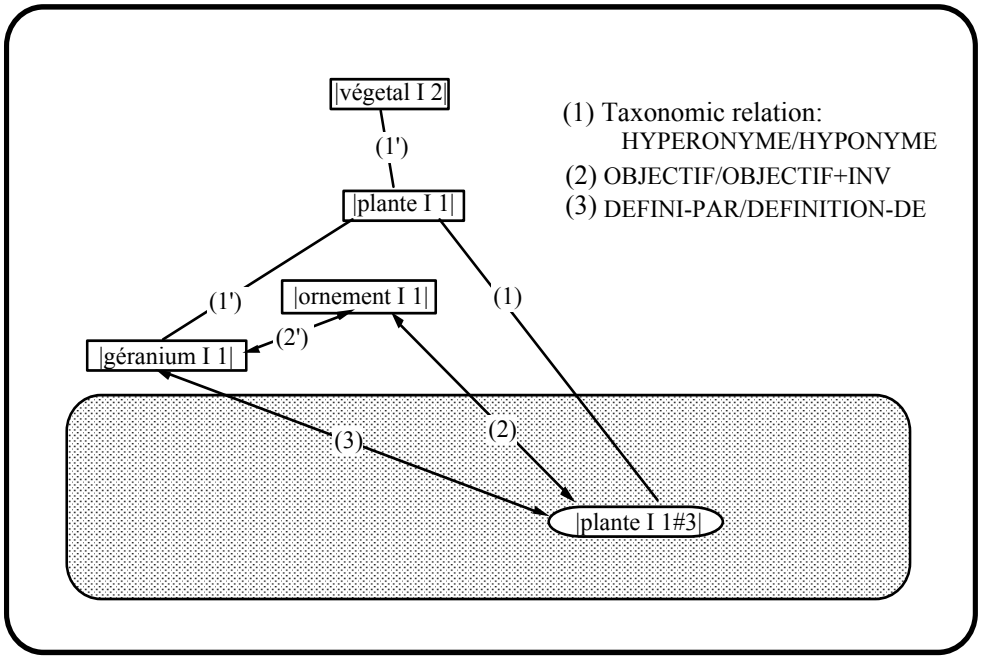

Fig. 5. Relational view of the concept |géranium I 1| (in the THESAURUS net).

(English gloss: DEFINI-PAR=defined by, DEFINITION-DE=definition of, géranium=geranium, ornement=ornament, plante=plant, végetal=vegetable)

[1] This could be very interesting also, in the opinion of the authors, in a multilingual environment: it is possible that, in another language, the concept equivalent to that which has been represented by the phrasal concept |pièce I 1#2| has its own significant, a word that denotes it. In this case, the phrasal concept based representation may be useful to represent the equivalence between both concepts.

[2] By means of rules fired following a forward chaining strategy.

[3] Lexical ambiguity comes from the definitions themselves; syntactical ambiguity is due mainly to the analysis process.

[4] The result of the transitivity rule that will be created will be the deduction of values for the R_3 attribute. The triples are stored in a facet of R_3.

[5] Considering only the set of triples declared until now.