An Intelligent Dictionary Help System

E. Agirre
X. Arregi
X. Artola
A. Díaz de Ilarraza
University of the Basque Country, Donostia, The Basque Country, Spain

F. Evrard
ENSEEIHT, Toulouse, France

K. Sarasola
A. Soroa
University of the Basque Country, Donostia, The Basque Country, Spain

INTRODUCTION

The Intelligent Dictionary Help System (IDHS) is a monolingual (explanatory) dictionary system.[1–4] Its design was conceived from the study of questions that human users would like to have 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 the deduction of implicit knowledge from the explicit dictionary information. The 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. Definitions have been analyzed 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, several enrichment processes have been executed on the DKB 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 resulting from 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.

This article provides an overview of IDHS, presents the process of construction of the DKB, and describes the knowledge representation model and the enrichment mechanisms. It also describes some inferential aspects of the system, presents some figures about the size and contents of the prototype built, and outlines some perspectives and derived works undertaken to deal with multilingual dictionary help environments. Finally, some conclusions are presented.

THE IDHS DICTIONARY SYSTEM

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

- The *dictionary knowledge base*, which represents the knowledge extracted from the dictionary by means of frame structures. It has been organized in different submodules and is explained in more detail below.

Encyclopedia of Library and Information Science
DOI: 10.1081/E-ELIS 120008829

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

- The *inference module*, which facilitates the inferencing capabilities of the system. The basic functionality is part of this module. More precise explanations are given below.
- The *communication module*, which interprets the questions posed by the user and translates them to the internal representation, and 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 interface module is not the focus of the work presented here. Fig. 1 shows the general architecture of IDHS.

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. Some of the functions implemented include definition queries, a search of alternative definitions, differences, relations and analogies between concepts, a thesauruslike word search, and verification of concept properties and interconceptual relationships.[8,9]

For instance, a definition request, Demande de Définition, takes as input a concept, an explanatory level, a dictionary, and a language, giving as output a definition with different levels of explanation: textual (the result is just the text associated to that definition), local (the answer gives the networklike representation of the textual definition), and inherited (it produces the networklike 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 (Le Plus Petit Larousse) French dictionary, but the requested explanatory levels are different: *textual* in the first example, *local* in the second one, and *inherited* in the third (see fixgraphic below).

```
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.
```

Fig. 1 General architecture of IDHS.

The next example shows the results of the thesaurus-like search of concepts, Recherche thésaurique. 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: (see fixgraphic below)

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 the KEE (knowledge engineering environment).

In that which concerns the construction of the system, all the knowledge represented in IDHS has been acquired from a conventional dictionary by means of parsing dictionary definitions using Natural Language Processing techniques. Two different steps were distinguished when building the DKB. First is the extraction of the information from the dictionary and its recording into a relational database—the DDB. This DDB was the starting point to create the object-oriented DKB, in step 2 (Fig. 2) that is the support of our deduction system.

Focusing on step 2 (construction of the DKB from the DDB), two phases are distinguished. First, information contained in the DDB is used to produce an initial DKB. General information about the entries obtained from the DDB [part of speech (POS), usage, examples, etc.] is conventionally represented—attribute-value pairs in the frame structure—whereas the semantic component of the dictionary (i.e., the definition sentences) has been analyzed and represented as an interrelated set of concepts. In this stage, the relations established between concepts could still in some cases be of lexical-syntactic nature. In a second phase, the semantic knowledge acquisition process is completed using the relations established in the initial DKB. This phase performs lexical and syntactical disambiguation, showing that semantic knowledge about hierarchical relations between concepts can be determinant for this.

BUILDING THE DKB

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

The dictionary was recorded in a relational database (the DDB). This DDB is the basis of every empirical

```
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.
```

Fig. 2 From the MRD to the DKB.

study that has been developed 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). The example below shows two different entries and the information associated in the database to their definition words, once tagging and disambiguation have been performed [in each example, (a) stands for definition texts, (b) for canonical forms, (c) for POS, (d) for orthomorphological alterations, and (e) are the English glosses].

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

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 the following three types of treatment:

1. Treatment associated with 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.
2. Treatment associated with synonymic definitions. In this case, an attribute representing the synonymic relation is used.
3. Treatment associated with definitions with a specific formula (specific relators). Different kinds of attributes are defined to represent the information conveyed by the formula.

The lexical-semantic relations between different concepts extracted from the analysis of the source dictionary are divided into paradigmatic (synonymy and antonymy, hypernymy/hyponymy) and syntagmatic relations (derivation, attributive, etc.).

REPRESENTATION OF THE DICTIONARY KNOWLEDGE: THE DKB

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

- KB-THESAURUS is the representation of the dictionary as a semantic network of frames, in which each frame represents an *one-word concept* (word sense) or a *phrasal concept*. Phrasal concepts represent phrase structures associated with 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 realized by means of metalinguistic relators, casuals, and so on. Other slots contain phrasal, metalinguistic, 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 metaknowledge 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, and so on. Units in KB-THESAURUS and KB-DICTIONARY are subclasses or instances of classes defined in KB-STRUCTURES.

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

When a dictionary entry such as *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* (because *cuiller* is a hyponym of *ustensile*) and so it will inherit some of the characteristics of *ustensile* (depending on the inheritance role of each attribute). The phrasal concept unit representing the noun phrase *cuiller plate* is treated as a hyponym of its nuclear concept (*cuiller I 1*).

KB-STRUCTURES: The Metaknowledge

This KB reflects the hierarchical organization of the knowledge included in the DKB. We focus on the LKB-STRUCTURES class, which defines the data types used in KB-DICTIONARY and KB-THESAURUS, and which organizes the units belonging to these KBs into a taxonomy.

Slots defined in LKB-STRUCTURES have associated aspects, such as the value class and the inheritance role, determining how values in children's slots are calculated. 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, whereas 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, which groups dictionary entries belonging to KB-DICTIONARY
- DEFINITIONS, which 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, which 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] in fact, this class is 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, which classify different types of concepts.
- PHRASAL-CONCEPTS is a class that includes concepts quite corresponding to Quillian's ''tokens''—occurrences of type concepts in the definition sentences. Phrasal concepts constitute 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, etc.), 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, subclass of the type concept |plante I 1| (see the example below), 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|), 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 comprised of the following two levels:

- *Definitory level*, in which the surface representation of the definition of each sense is made. Such morpho-syntactic features as verb mode, time, and determination 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*, which reflects the relational view of the lexicon. It supports the deductive behavior of the

system and is made up of *relational attributes*, which may eventually contain deduced knowledge. These attributes are defined in the class TYPE-CONCEPTS and implement the interconceptual relations: ANTONYMES, AGENT, CARACTERISTIQUE, SORTE-DE, CE-QUI, and so on.

KB-DICTIONARY: From Words to Concepts

This KB contains the links between each dictionary entry and its senses.

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 to obtain 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: one which corresponds to the *definiendum*, and the phrasal concept, which represents the noun phrase of the definition. Moreover, the units that represent *plante* and *ornement* are also to be created (if they have not been previously created because of 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 such aspects as determination and genre, 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 the case of the facet OBJECTIF in the slot DE in |plante I 1#3| below, which states that the slot value will be probably promoted to the OBJECTIF—purpose, goal—relationship.)

The following shows 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).

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

It is interesting to note that a relation of *conceptual equivalence* is established between |géranium I 1| and |plante I 1#3| because these units actually represent the same concept (|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 now give another example, the case of two definitions stated by means of two different stereotyped formulae belonging to the lexicographic metalanguage. 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 metalanguge 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|. In the case of the definition of *ajusteur I 1*, however, 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

Phrasal concepts frequently represent ''unlabeled'' 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* or a noun meaning *pièce de métal*. This is not the case of the phrasal concepts that are linked to type concepts by means of the relation DEFINI-PAR/DEFINITION-DE, however, because there the phrasal concept is another representation of the concept being defined. (See the example of the definition of *géranium I 1* above.) In the representation model proposed in this article, phrasal concepts denote concepts that are typically expressed in a periphrastic way and that do not necessarily have any corresponding entry in the dictionary.[a]

[a]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.

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 us, in fact, a virtual set of usage examples that may be useful for different functions of the final system.

ENRICHMENT PROCESSES PERFORMED ON THE DKB

In this section, the enrichment processes accomplished on the DKB are explained. Two phases are distinguished: 1) the enrichment obtained during the construction of the initial DKB, and 2) where different tasks concerning mainly the exploitation of the properties of synonymy and taxonymy have been performed.

Enrichment Obtained During Construction of 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 were 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.

Second Phase in Enrichment of the DKB

Several processes have been carried out to infer new facts to be asserted in the DKB, by means of rules fired following a forward-chaining strategy. The enrichment obtained in this phase concerns the following two 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 disambiguation. The use of the lexical-semantic knowledge about hierarchical relations contained in the DKB can be determinant to reduce the level of lexical and syntactical ambiguity.[b] 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.

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 that 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

Following, some aspects concerning the second point are discussed.

In IDHS, the relationships among the different lexical-semantic relations can be 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 $(R_1R_2R_3)$, where R_i represents a lexical relation.[c] 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:

$$\text{if } X\ R_1\ Y \text{ and } Y\ R_2\ Z \text{ then } X\ R_3\ Z$$

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

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

When the value(s) of the attribute R_3 is (are) asked, a reading demon (attached to the attribute) creates the rule and fires the reasoning process under 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[11,12] can be easily expressed by stating the triple (PARTIE-DE PARTIE-DE PARTIE-DE), and 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 the following:

- Combination of meronymic and nonmeronymic relations:
  (PARTIE-DE LOCATIF LOCATIF)
  (LOCATIF HYPERONYME LOCATIF)
  (MEMBRE-DE HYPERONYME MEMBRE-DE)
- Combination of relations derived from the definition metalanguage:
  (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 fired following the same reasoning strategy.

The following is the rule derived from the last triple along with 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 purpose of X is Y (entity)
Y CE-QUI Z **;;;** Y ''est ce qui'' Z (action)
then X OBJECTIF Z **;;;** the purpose 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|

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 pretty 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, 19,691 relational arcs—interconceptual relationships—had been established. After the enrichment processes, the number of relational links have been incremented up to 21,800 (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%.[d]

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 rate given by the size of the sample).

Concerning the deduction of semantic knowledge, two considerations arise. First, the use of dubious lexical rules, such as the transitivity of synonymy, has led to some errors in the prototype. Second, 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. We are now using a knowledge-based technique for lexical disambiguation of free-running text[13] and applying it to dictionary definitions.

PERSPECTIVES

A Multilingual Dictionary Help System

Currently, a multilingual environment is being designed on the basis of different dictionaries. The MLDS (multilingual dictionary system, an extension of IDHS) is conceived as an intelligent help system for human translators,[14–16] where two monolingual dictionaries (French and Basque) constitute the KB along with a bilingual dictionary that establishes equivalence-links among concepts from the monolingual dictionaries. This allows the system to enrich its functionality, as is shown next.

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

There are some well-known problems with lexical gaps when 1) there is no single word in the target language to

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

```
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".
```

express the source concept, which can be solved giving *phrasal concept equivalents*, and 2) the source concept does not appear as an entry in the bilingual dictionaries. In this case, 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 above, 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 (see fixgraphic above).

Intelligent Dictionaries as Lexical Information Sources

The problem of querying very diverse sources of lexical information—lexical and dictionary databases, heterogeneously structured electronic dictionaries, or even language processing programs such as lemmatizers or part-of-speech taggers—using for that a unique and common query language is addressed in Ref. [17] from the field of information integration. This is done by building a federation that integrates various lexical resources, without forcing us to convert them into a single and standard representation schema. In Ref. [17], a general conceptual model for describing lexical knowledge is presented, as well as the way to describe each source in terms of the classes and relationships of the general model. The so-called *local-as-view* paradigm is used for describing each lexical source as a view over the general conceptual model. Both the conceptual model and the sources have been described and implemented using a description logic language, and an algorithm that translates queries from the general model into each particular source schema has also been implemented.

The lexical resources integrated in such a federation can be accessed by means of a common query language based on the general conceptual model. We are now working in describing in such a way different lexical resources, including the intelligent dictionary help systems depicted in this article.

CONCLUSION

The general objective of IDHS is to assist a human user in language comprehension or production tasks. The system provides a set of functions that allow the user to query the dictionary and to obtain from it both explicit and implicit knowledge.

Moving from the monolingual to the multilingual environment, IDHS has been used in the design and implementation of a computerized translation-oriented dictionary (MLDS) that helps human translators in choosing suitable target lexical units that correspond with those that are in the source text. A new lexical KB 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.

Both IDHS and MLDS will be integrated into a Federation of Heterogeneous Lexical Databases that also includes more conventional lexical databases and dictionaries, thus constituting a large lexical information store. This lexical bank will be accessible by means of a unique and common query language.

In that what concerns more specifically to IDHS, a methodology for the extraction of semantic knowledge from a conventional dictionary is described in the article. This extraction was founded on a systematic study of dictionary definitions. As a result of this study, the characterization of the different lexical-semantic relations between senses—which is the basis for the proposed DKB representation schema—was established.

A frame-based knowledge representation model was described and used in the intelligent dictionary help system to represent the lexical knowledge acquired automatically from a conventional dictionary. The characterization 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.

REFERENCES

1. Artola, X. *HIZTSUA: Hiztegi-Sistema Urgazle Adimendunaren Sorkuntza eta Eraikuntza/Conception d'un Système Intelligent d'Aide Dictionnariale (SIAD)*; University of the Basque Country (UPV-EHU): Donostia, 1993, Ph.D. Thesis.
2. Artola, X.; Evrard, F. Dictionnaire Intelligent d'Aide à la Compréhension. In *Actas IV Congreso Internacional EURALEX, Benalmádena (Spain), 1990*; Biblograph: Barcelona, 1992; 45–57.
3. Agirre, E.; Arregi, X.; Artola, X.; Díaz de Ilarraza, A.; Sarasola, K.; Soroa, A. Constructing an intelligent dictionary help system. Nat. Lang. Eng. **1996**, *2* (3), 229–252.
4. Agirre, E.; Arregi, X.; Artola, X.; Díaz de Ilarraza, A.; Evrard, F.; Sarasola, K. Intelligent Dictionary Help System. In *Applications and Implications of Current Language for Special Purposes Research*; Fagbokforlaget: Bergen, 1994; Vol. I, 174–183.
5. Alshawi, H. Analysing Dictionary Definitions. In *Computational Lexicography for Natural Language Processing*; Boguraev, B., Briscoe, T., Eds.; Longman: New York, 1989; 153–169.
6. Agirre, E.; Arregi, X.; Artola, X.; Díaz de Ilarraza, A.; Evrard, F.; Sarasola, K. Lexical Knowledge Representation in an Intelligent Dictionary Help System. In *Proceedings of COLING'94, Kyoto (Japan)*; 1994; 544–550.
7. Agirre, E.; Arregi, X.; Artola, X.; Díaz de Ilarraza, A.; Evrard, F.; Sarasola, K. A Methodology for the Extraction of Semantic Knowledge from Dictionaries Using Phrasal Patterns. In *Proceedings of IBERAMIA'94, Caracas (Venezuela)*; 1994; 263–270.
8. Arregi, X.; Artola, X.; Díaz de Ilarraza, A.; Evrard, F.; Sarasola, K. Aproximación Funcional a DIAC: Diccionario Inteligente de Ayuda a la Comprensión. In *Proceedings of SEPLN 11*; SEPLN, 1991; 127–138.
9. Agirre, E.; Arregi, X.; Artola, X.; Díaz de Ilarraza, A.; Evrard, F.; Sarasola, K. IDHS, MLDS: Towards Dictionary Help Systems for Human Users. In *Semantics and Pragmatics of Natural Language: Logical and Computational Aspects*; Korta, K., Larrazabal, J.M., Eds.; ILCLI Series, Donostia, The Basque Country 1995; Vol. 1, 167–188.
10. Quillian, M.R. Semantic Memory. In *Semantic Information Processing*; Minsky, M., Ed.; MIT Press: Cambridge, MA, 1968; 227–270.
11. Cruse, D.A. *Lexical Semantics*; Cambridge University Press: Cambridge, 1986.
12. Winston, M.E.; Chaffin, R.; Herrmann, D. A taxonomy of part–whole relations. Cogn. Sci. **1987**, *11*, 417–444.
13. Agirre, E.; Rigau, G. Word Sense Disambiguation Using Conceptual Density. In *Proceedings of COLING'96, Copenhagen (Denmark)*; 1996.
14. Arregi, X. *ANHITZ: Itzulpenean Laguntzeko Hiztegi-Sistema Eleanitza/ANHITZ: Multilingual Dictionary Help System for Translation Tasks*; University of the Basque Country (UPV-EHU): Donostia, 1995, Ph.D. Thesis.
15. Agirre, E.; Arregi, X.; Artola, X.; Díaz de Ilarraza, A.; Patel, H.; Sarasola, K.; Soroa, A. A Computerised Translation-Oriented Dictionary. In *Proceedings of NLP + IA/TAL + AI 96, Moncton (Canada)*; 1996.
16. Agirre, E.; Arregi, X.; Artola, X.; Díaz de Ilarraza, A.; Sarasola, K.; Soroa, A. MLDS: A Translator-Oriented MultiLingual Dictionary System. Nat. Lang. Eng. **1999**, *5* (4), 325–353.
17. Artola, X.; Soroa, A. An Architecture for a Federation of Highly Heterogeneous Lexical Information Sources. In *Proceedings of the IRCS Workshop on Linguistic Databases, Philadelphia (USA)*; 2001; 17–23.

FURTHER READING

Amsler, R.A. A. Taxonomy for English Nouns and Verbs. In *Proceedings of the 19th Annual Meeting of ACL*; ACL, 1981; 133–138.

Arango Gaviría, G. *Une Approche Pour Amorcer le Processus de Compréhension et d'Utilisation du sens des mots en Langage Naturel*; Publications du Groupe de Recherche

Claude François Picard: Paris, 1983. Thèse de 3ème cycle (Univ. Paris VI).

Byrd, R.J.; Calzolari, N.; Chodorow, M.S.; Klavans, J.L.; Neff, M.S.; Rizk, O.A. Tools and methods for computational lexicography. Comput. Linguist. **1987**, *13* (3,4), 219–240.

Calzolari, N.; Picchi, E. Acquisition of Semantic Information from an On-line Dictionary. In *Proceedings of COLING'88, Budapest*; 1988; 87–92.

Chodorow, M.S.; Byrd, R.J. Extracting Semantic Hierarchies from a Large On-line Dictionary. In *Proceedings of ACL*; ACL, 1985; 299–304.

Chouraqui, E.; Godbert, E. Représentation des Descriptions Définies dans un Réseau Sémantique. In *Actes 7ème Congrès Reconnaissance des Formes et Intelligence Artificielle (AFCET-INRIA), Paris*; 1989; 855–868.

Computational Lexicography for Natural Language Processing; Boguraev, B., Briscoe, T., Eds.; Longman: New York, 1989.

Copestake, A. An Approach to Building the Hierarchical Element of a Lexical Knowledge Base from a Machine-Readable Dictionary. In *Paper Read at the First Int. Workshop on Inheritance in NLP, Tilburg (The Netherlands)*; 1990.

Litkowsky, K.C. Models of the semantic structure of dictionaries. Am. J. Comput. Linguist. **1978**, *81*, 25–74.

Markowitz, J.; Ahlswede, T.; Evens, M. Semantically Significant Patterns in Dictionary Definitions. In *Proceedings of the 24th Annual Meeting of ACL, New York*; 1986; 112–119.

McRoy, S. Using multiple knowledge sources for word sense discrimination. Comput. Linguist. **1992**, *18* (1).

Pazienza, M.T.; Velardi, P. A Structured Representation of Word-Senses for Semantic Analysis. In *Proceedings of the 3rd European Conference of ACL, Copenhaguen*; 1987; 249–257.

Tsurumaru, H.; Hitaka, T.; Yoshida, S. An Attempt to Automatic Thesaurus Construction from an Ordinary Japanese Language Dictionary. In *Proceedings of COLING'86, Bonn*; 1986; 445–447.

van den Hurk, I.; Meijs, W. The dictionary as a corpus: Analyzing LDOCE's definition-language. Corpus Linguist. II **1986**, 99–125.

Vossen, P.; Meijs, W.; den Broeder, M. Meaning and Structure in Dictionary Definitions. In *Computational Lexicography for Natural Language Processing*; Boguraev, B., Briscoe, T., Eds.; Longman: New York, 1989; 171–192.

Wilks, Y.; Fass, D.; Cheng-Ming, G.; McDonald, J.E.; Plate, T.; Slator, B.M. Providing machine tractable dictionary tools. Mach. Transl. **1990**, *5*, 99–154.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016