Towards a Development Methodology for Managing Linguistic Knowledge Bases

F. Sáenz and A. Vaquero
Departamento de Sistemas Informáticos y Programación, Universidad Complutense de Madrid, E-28040 Madrid, Spain

{fernan,vaquero}@sip.ucm.es

Abstract

We are on the way of defining a methodology aimed to create software tools supporting linguistic knowledge bases. One of our main concerns is to formally represent knowledge using a sound software engineering approach. In this setting, we first consider the linguistic concepts found in a multilingual dictionary, as vocabulary, meanings, semantic categories, semantic relationships, and (tree-shaped) taxonomy. Next, we start on ontologies, considering concepts as syntactic categories, orthography, phonology, syntactic features, lexical semantics and relations, and so on. We have represented in the conceptual levels these concepts, by using the well known entity-relationship model. In addition, we have applied the design cycle of databases in order to also fulfill the logical and physical models for representing the linguistic concepts at each development stage. In addition, we have developed both authoring and querying tools for both stages, and a migration procedure for interfacing them. Taking into account that the different existing linguistic knowledge bases have been built without following any formal methodology, our approach adds a way of integrating linguistic resources with a common structure. The resulting framework is useful for several applications, including multilingual information retrieval, document classification, and language translation, and also for their exploitation in education.

1. Introduction

Lack of standardisation is broadly felt as a very undesirable state into the community around ontologies, lexicons, and so on. For instance, standard terminology for a common reference ontology is a goal to be reached. But attention has not yet been paid on subjects about development methodologies for building the software tools supporting and handling those types of knowledge bases. We claim for this aspect of methodology as necessary in order to integrate the diverse available information systems of this kind now and in the future. A more or less automated incorporation of lexical and ontological databases into a

common information system requires compatible software architectures and sound data management from the different databases to be integrated. With this vision in mind, paying attention to the software engineering aspects along the development of these kinds of systems from the beginning is mandatory.

In this paper, we present our ongoing work on developing sound conceptual models for terminological and ontological databases, with the aim of developing tools which can manage such lexical and semantic resources. There are many reasons for developing such tools. For instance, lack of the kind of dictionaries we propose (as will be introduced later) has been felt, as [2] states: "... we imagine, for some distant future, an online lexical resource, which we can refer to as a 'frame-based' dictionary, which will be adequate to our aims. In such a dictionary (housed on a workstation with multiple windowing capabilities), individual word senses, relationships among the senses of the polysemic words, and relationships between (senses of) semantically related words will be linked with the cognitive structures (or 'frames'), knowledge of which is presupposed by the concepts encoded by the words." In addition, it is well known the applications of ontological resources for different fields, as language translation, information retrieval, document summarisation, document classification, software localisation, language teaching, and so on.

Subjects about electronic dictionaries for diverse natural language processing applications have been extensively studied [12], as well as Lexical Databases [7], World Knowledge Bases [4], ontologies [6], and the like. But there are no references on how these information systems have been built, and generally, there is no registered information about how they have been developed and upgraded along their life. Moreover, tools for managing ontology-based information systems have been described [8], but there is neither formal support for their conceptual models nor a software engineering approach for the development. Our tools do enjoy from these two important issues. We have followed the classical relational database design (based on the conceptual, logical, and physical models) and software engineering techniques (based on UML).

The rest of the paper is organised as follows. Section 2 presents some concepts which has to be embodied in the lexical and ontological resources for their relevance in building different applications. The next three sections present the different conceptual models we present for several linguistic resources. For all of them, we have followed a classical relational database design cycle. First, from the conceptual model of each linguistic resource, we have developed the entity relationship model. Second, in the logical design stage, we have developed the relational model. Finally, in the physical design stage, we have developed the physical database schema. Section 3 presents the first conceptual model we develop to build a bilingual dictionary and that embodies some of the concepts listed in Section 2. Section 4 presents an extension of the first conceptual model in order to achieve a (dynamic) multilingual language. Section 5 develops a conceptual model for an ontology (we have selected MikroKosmos [6]). Section 6 sketches some tools we have developed for querying and building dictionaries, building ontologies and lexicons, and migrating information from our electronic

dictionary to MikroKosmos. Finally, Section 7 summarises our conclusions and points out some future work.

2. Concepts to be Attained

In this section, linguistic concepts incorporated in computing systems devoted to natural language processing are pointed out because of their relevance in the definition of the conceptual models.

2.1 Order, Classification, and Ontology

Typically, monolingual dictionaries show an alphabetical order that can be seen as a simple term classification: terms are classified in singletons by its lexicographic form. Other possible less naïve classifications are derivative (root-shape), grammatical, and semantic. Derivative classifications [5] are not common, and grammatical classifications are not intended for dictionaries. Finally, semantic classification groups terms by semantic categories (for instance, synonym and antonym dictionaries, or ideological dictionaries [1].) Semantic categories not also allow meaning classification, but the more meaningful taxonomy of meanings. Conventional lexical databases, such as WordNet [7], have term classification such as synonymy (grouped in the so called synsets.) Ontologies go beyond by playing the role of meaning taxonomy [9]. Our tools do support this important concept as will be explained along the paper.

Semantic categories are useless for term lookups since meanings will correspond, in general, to a set of (synonym) terms[1]. However, it has an important role on learning by both using and authoring dictionaries because each meaning of a given term (polysemy and/or homonymy) is precisely identified by its semantic category (categories from now on, for the sake of brevity), instead of the usual nonsense sequential number[2]. Therefore, semantic categories provide classification for meanings, and such classification can be arranged in a taxonomy. But this does not straightforwardly imply a term order since meanings are abstract ideas that cannot be expressed in general by one distinctive word[3]. It is commonly acknowledged that the best order for lookups is lexicographic (a derivative classification is a counterexample for this, but it still keeps a lexicographical order by repeating entries and adding links.) Figure 1 resumes the order for taxonomies in a hierarchy; it shows a taxonomy of categories along with the set of terms belonging to each category. From this point of view, there is a complete lexicographic order (provided categories are identified with terms or phrases.) A hierarchy is a natural structure for meaning classification. Each node in the hierarchy corresponds to a category. In principle, every category in the hierarchy can be used, no matter its

---

[1] Nevertheless, there are other kinds of term lookups as ideological dictionaries show.

[2] However, meaning identifications by numbers also show a coarse classification; e.g. Tech. for Technical.

[3] The question is: Which is the best word to represent a meaning? In general, there are several (synonym) words representing the same meaning.

hierarchy level. It must be noted that every category in the hierarchy contains at least the term which names the category, so that all categories are non-empty. On the other hand, the creation of new categories as intersection of several predefined ones should be avoided, in order to reach compactness.

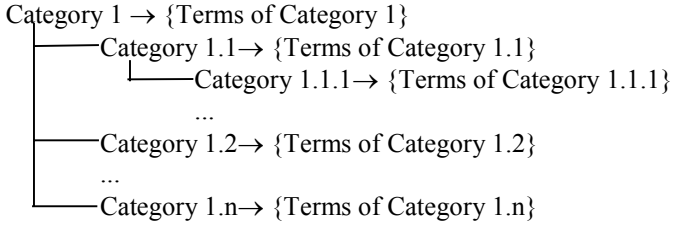

Figure 1. A Taxonomy

There are a number of advantages in classifying meanings as a taxonomy. First, meaning taxonomy is a useful facility for an electronic dictionary because meanings embody additional semantics which provides more information to the reader (more than that of sequential numbers noted above.) Second, the system may also gain a new dimension because it is possible to automatically generate specialised dictionaries under different categories (a sports dictionary may deal with soccer, tennis, or baseball dictionaries.) Third, it helps to develop a balanced dictionary by adding enough terms from different categories. Having the terms classified, it is easy to check out how many terms are under a given category. Fourth, it also helps to distribute the work between several authors by assigning categories to authors. A team of authors may develop a complete specialised dictionary by dividing the work by categories so that collaborative work is promoted for students.

2.2 Polysemy and Synonymy

In every language there exists the well known naming problem [3], which consists of two elements: one is polysemy (under the synchronic point of view, that is, embodying polysemy itself and homonymy), by which a term can have several meanings; and the other is synonymy, by which one meaning can be assigned to different terms, as can be observed in Figure 2. In this Figure, Term 1 and Term 2 are synonyms and have a shared meaning, as so for Term 2 and Term3, under another meaning. Moreover, Term 2 is polysemic since it has two possible meanings.

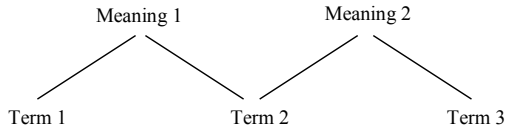

Figure 2. Polysemy and Synonymy

2.3 Relationships

2.3.1 Basic Relationships

Here we do some remarks about the relationships between categories, meanings and terms. On the one hand, a given term can belong to several categories under different meanings. On the other hand, a given term can belong to several categories under the same meaning. Figure 3 shows two categories (C1 and C2) which respectively contain the meanings {M11, M12, M} and {M, M21, M22}. Each meaning has one or more terms associated. The term T2 is associated to meanings M12 and M21, which respectively belong to categories C1 and C2. We also show the term T that is assigned to meaning M, which belongs to both categories C1 and C2. Polysemy is present in T2, and synonymy is also present in T3, and T4, as it can be seen. T1 is neither polysemic nor synonym. TC1 and TC2 are the terms used to denote categories C1 and C2, respectively.

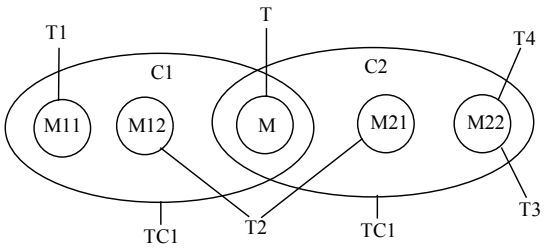

Figure 3. Relationships among categories, meanings and terms. Extensional definition

In this figure, the set of meanings {M11, M12, M} in C1 is the extensional definition of category C1. We must also note that a category has a meaning described by a definition. This figure does not embody this fact. In order to embody the meanings related to categories, we transform the scheme of Figure 3 to the one depicted in Figure 4. Now, C1 is the meaning of the category C1, and TC1 is the term assigned to such meaning, and the same applies to C2 and TC2. Then, we have one more meaning in each category. This meaning is the intensional definition of the category.

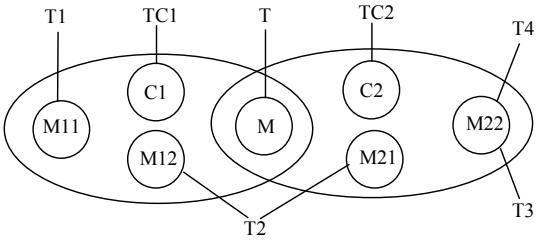

Figure 4. Relationships among categories, meanings and terms. Intensional definition

For a given language, we have a set of terms that holds the relationships with categories and meanings shown in Figure 4. If we now think of several languages, the same applies for each one. Then, relationships between terms from different languages come from considering jointly the involved schemes .

2.3.2 Other Lexical and Semantic Relationships

For all languages, knowledge in the discourse universe belongs to two types: conceptual and linguistic. Terms and sentences refer to concepts, but they have particular structural and morphological features for each language. Language mastery includes the ability to distinguish both knowledge types. In fact, language mastery traverses several stages until it is learned how to distinguish between concepts and the linguistic way of expressing them in a given language. It is needed to learn concepts and their relations, lexicon and linguistic properties of terms, compositionality defined by the syntactic structure and links between terms and concepts. These goals are relevant also for pedagogical interests.

Although ontology is not exactly the same as conceptual knowledge of discourse, there is no computer mean more adequate for representing it. All of the relations (meronimy, holonimy, hiperminia, hiponimia, and so on) represented in the more complete lexical databases as WordNet, are in ontology-based databases, as the MikroKosmos system, which is based in the ontology Ontos; but in these cases, relations are present in a level-structured way. In an ontology, concepts and their relations are represented, whereas each lexicon has the terms for each language and their linguistic properties, as well as their mappings with ontology concepts. The mapping between ontology and lexicons is the key for successfully coordinate all of the lexical and semantic relations.

3. Conceptual Model of the Terminological Database for a Bilingual Dictionary

Our work in developing the tools is based on a sound conceptual model for the terminological database (TDB) which shall eventually hold the terms, definitions, meanings, and semantic categories. Since it is intended to deal with two or more languages (bilingual or multilingual dictionaries), we need to represent instances of

terms, textual definitions, and textual semantic categories for each language, but, as meanings are not language dependent, we shall use unique representations for them.

The entity-relationship model is used to describe the conceptual model we propose, shown in Figure 5. In this figure (following some recommendations in [10,11]), entity sets are represented with rectangles, attributes with ellipses, and relationship sets with diamonds connecting entity sets with undirected lines (many to many mapping cardinality). A one to many mapping cardinality from entity set A to entity set B is represented by an arc from B to A. (There is no such relationship set type in this first model.) Undirected lines also connect attributes to entity sets. Relationship set and entity set names label each diamond and box, respectively.

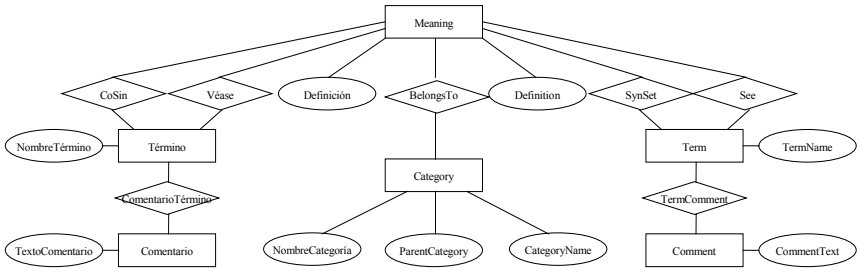

Figure 5. Entity-Relationship Model for an English-Spanish TDB

In this figure we show an instance of a bilingual terminological database for Spanish and English languages; further, its extension to support several languages is presented in Section 4. In the following, we firstly describe entity sets, then relationship sets, and, finally, attributes.

The entity set Meaning is the central entity set other entity sets rest on. In fact, this is the entity set which is language independent. The entity set Term represents all the English terms that compose the terminological database. The entity set Category denotes the category each meaning belongs to. The entity set Comment represents the comments about each term.

The relationship set SynSet between Meaning and Term denotes the English synonym set and it is many to many since a synonym set contains several terms, and a term may be contained in several synonym sets (obviously, with different meanings.) The relationship set See denotes the set of English terms related under a given meaning. This relationship which connects Meaning and Term is many to many because a meaning may refer to several English terms, and one term may be referenced by several meanings. The relationship set BelongsTo between Category and Meaning is many to many since many meanings are in a category, and a meaning could be in several categories (this situation is expected to be reduced to the minimum since the goal is to keep the classification as disjoint as possible). This relationship set embodies the fact that our classification is not lexical (there is not a direct relationship between Category and Term) but semantic (we relate meanings to categories, i.e., we categorise meanings.) The relationship set

TermComment is many to many since a term may have several comments attached and a comment may refer to several terms.

The entity set Category has three attributes: CategoryName, NombreCategoría, and ParentCategory. The first two correspond to the textual name of the category in each considered language, English and Spanish, respectively. The last attribute, ParentCategory, represents the links in the taxonomy by relating a category with its parent. Since each entity Category has a monovalued attribute for parent, this means that we restrict taxonomies to trees. If we change this attribute by a multivalued attribute (or, alternatively, we connect the entity set Category with itself via a relationship set named ParentCategory), we allow a taxonomy graph instead of a tree. Meaning has two attributes: Definition and Definición, which correspond to the textual definition in the same considered languages. Term has one attribute: TermName, which denotes the textual term name. CommentText is an attribute which holds the textual comment for each term. The remaining entities and relationship sets (CoSin, Véase, Término, ComentarioTérmino y Comentario) are homologous to the ones in the other language (SynSet, See, Term, TermComment, and Comment.)

We have also developed (but not shown here for reasons of space) the logical and physical models for the development of our terminological database, which follow the design cycle of classical database design that ensures us a formal way of defining the data fundamentals the tools will adhere to.

4. Conceptual Model of the Terminological Database for a Multilingual Dictionary

We have developed a conceptual model the terminological database for a dynamic multilingual dictionary. With *dynamic* we refer to the user's ability for modifying the number of languages present in the dictionary without altering the database schema; in particular, the entity-relationship model. Figure 6 shows the model.

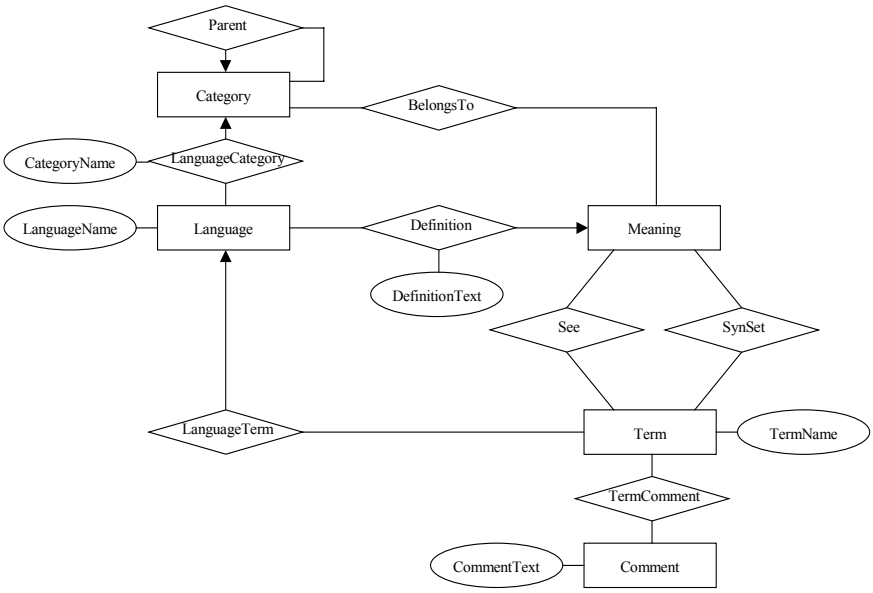

Figure 6. Entity-Relationship Model for a multilingual TDB

This model is more elaborated than the previous one in order to represent its dynamic feature.

A new entity is needed, Language, which denotes all the languages to be hold in the multilingual dictionary. The entity-relationship structure of meanings, terms, and comments is similar to the first conceptual model. However, the entity sets Term and Comment refer to all of the sets of terms and comments irrespective of the language. The key that indicates the language a term belongs to is the relationship set LanguageTerm.

On the one hand, by contrast, comments are linked directly to the terms by the relationship set TermComment. The comment itself is represented by the attribute CommentText of the entity set Comment.

On the other hand, the entity set Category is linked to a language via the relationship set LanguageCategory. In this case, we have to make explicit the language of a category since the category is independent from the language. Note that the attribute CategoryName is now linked with the relationship set LanguageCategory, that is, whereas the concept Category is independent from the language the text which describes the concept is not. Here, we have opted to use a relationship set Parent in order to avoid category hierarchies describing graphs (so, we have used a one to many mapping cardinality.) However, there is a lack of constraint information for describing trees in the conceptual model. For instance, we can represent forests with this model. Therefore, additional constraints are added in the conceptual stage as documentation (which has to be obeyed by the implementation).

Now, definitions are not attributes of the entity set Meaning; they are otherwise modelled with the attribute DefinitionText of the relationship set Definition, which link Meaning and Language (that is, a meaning has a definition in a given language.)

5. Conceptual Model of the Ontology for MikroKosmos

In order to be able of representing more detailed information about semantics and grammatical properties, we recourse to a database based on ontology. In this context, an ontology is a structured representation of world knowledge by means of symbols that represent the (language-independent) meanings, and possible relationships between them. The symbols are defined as concepts in the ontology, and also used to represent word meanings in lexicons.

Ontologies play an important role in NLP applications since they have an structure focused to the representation of knowledge about the world or a world domain. They hold symbols for meaning representation, organises these symbols in a tangled subsumption hierarchy, and interconnects these symbols using a rich system of semantic relations defined among the concepts. A concept is a primitive symbol for meaning representation with attributes and relationships with other concepts. An ontology is a network of such concepts.

We have selected [6] as an appropriate lexical database based on ontologies because of its structure. This structure is sufficient rich to support not only the conceptual and linguistic knowledge supported by the first tools previously described, but all the surplus required to improve the language mastery.

The ontology structure in [6] can be viewed as a directed graph with concepts as nodes. There are semantic relationships among nodes. The root concept is ALL (cfr. Figure 7) whose children are OBJECTS, EVENTS, and PROPERTIES.

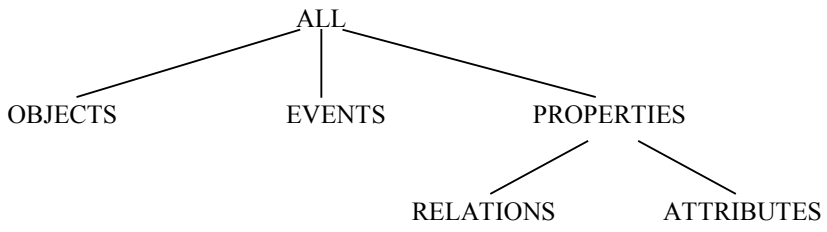

Figure 7. Ontology Hierarchy

One or more lexicons (for several languages) must be linked to the ontology in order to represent the language-dependent knowledge of the discourse. Lexicons are intended to hold terms and their lexical information. For instance, lexicons hold syntactic category, orthography (abbreviations and variants), phonology,

morphological irregular forms or class information, syntactic features such as attributive, indication of sentence-level syntactic inter-dependencies (including subcategorization), lexical semantics, meaning representation, lexical relations (e.g., collocations), pragmatics hooks (e.g. for deictics, and stylistic factors), and annotations (user, lexicographer, and administrative information, such as modification audit trail, example sentences, definition in English, etc.) Through the lexicon, the semantic information can be located for a given term. Note that there is semantic information in both the ontology and the lexicon so that language-neutral meanings are stored in the former, and language specific information in the latter.

Figure 8 shows the entity-relationship model for the MikroKosmos ontology (ONTOLOGY), together with the model for the lexicon (LEXICON) and the connections between them (LINK). This figure represent one ontology which can be connected to many lexicons belonging to different languages.

The entity set Concept represents the concepts in the ontology (this entity set is close to Meaning in the former conceptual models). The entity set Relation represents the different relations which may be defined among concepts in the ontology. The entity set Attribute represents the different attributes which may be attached to concepts in order to describe them. These two last entity sets stands for "types of"; the instance relations are represented by the relationship set RelCon, and the instance attributes by the relationship set AtrCon. Finally, Term is the entity set representing terms belonging to a lexicon. In fact, this entity set represents the set of all of the lexicons. Each lexicon can be distinguished by the set of all the instance terms so that they have the same value for the attribute Language.

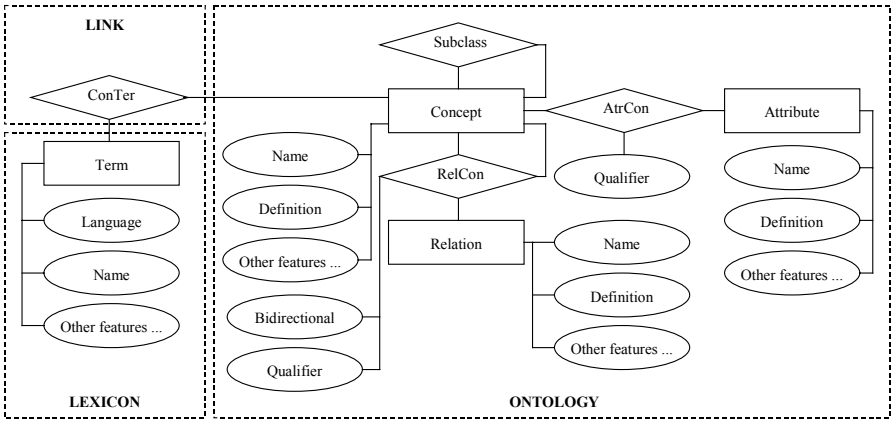

Figure 8. Entity-Relationship Model for the MikroKosmos Ontology

The relationship set Relation represents the link among two concepts by a relation. Each instance relation represents the link of one concept with another under a given relation type. Such relations are not bidirectional unless they are explicitly defined by the (boolean) attribute Bidirectional. The Qualifier attribute represents

additional information which drills down the instance; for instance, by adding a value that further makes concrete the relation. The relationship set Subclass represents the relation "is a" in the sense of objet oriented programming. Here, graphs are allowed to represent object containment. The link among the ontology and the lexicon is defined by the entity relationship set ConTer. This set contains all the pairs <Term, Concept> which define the concept each term represents. Note that this has a many to many mapping cardinality (polysemy and synonimy). The mapping cardinalities of the remaining relationship sets should be clear.

6. Tools for the Linguistic Resources

We have developed several tools for the above linguistic resources, namely: a tool for querying dictionaries (query tool), a tool for creating dictionaries (author tool), a tool for creating ontologies (ontology tool), a tool for creating lexicons (lexicon tool), and a tool for migrating data from a dictionary to an ontology-based information system (migration tool).

The querying tool is a query interface which allows the user to easily recover the information about both English and Spanish terms as well as their relationships from the terminological database. This database holds the terms, categories, their attributes, and the relationships. The interface allows the user to navigate the semantic categories, also allowing to retrieve the relevant information of any term (definition, other related terms, translation, synonyms, …). Figure 9 shows a snapshot of one screen of the interface.

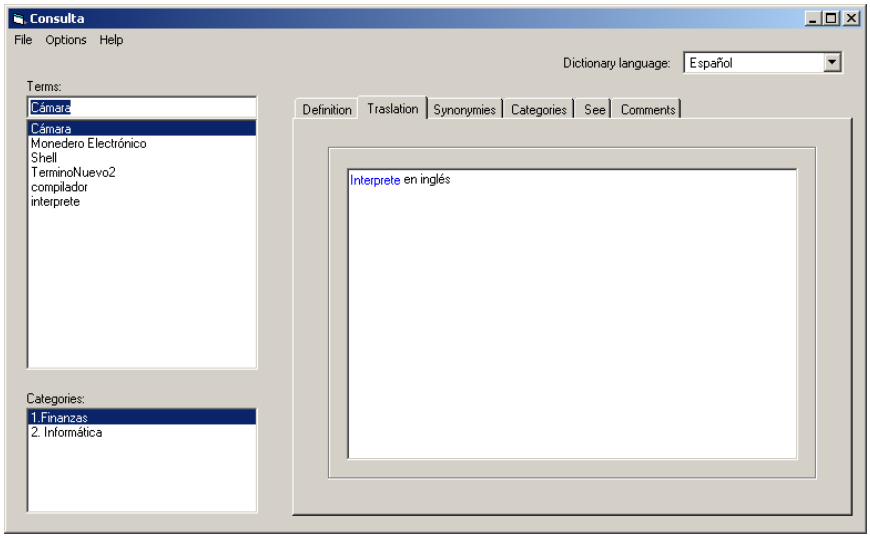

Figure 9. A Screen from the Querying Tool Interface

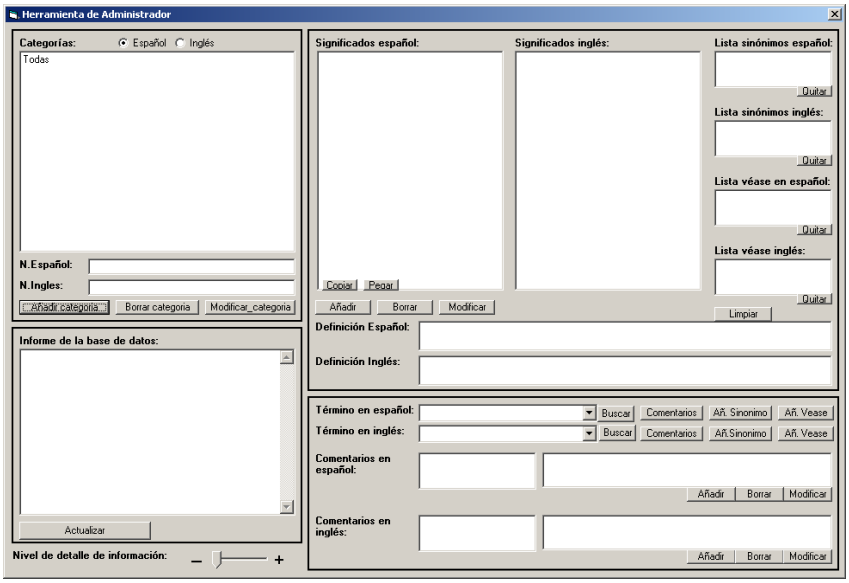

Figure 10. A Screen from the Author Tool Interface

The author tool allows the author to add new terms to the terminological database, and all the relevant information, such as its definition, semantic categories, meanings, synonym sets, and related terms. We have developed a Spanish user interface for this tool (easily rewritable for allowing to customise the use of any other language, as we have already done for the previous tool), and it consists mainly of one Author window. It has several areas for semantic category management, meaning management, synonyms and related terms management, and database consistency control. Figure 10 shows a snapshot of one screen of the interface.

The ontology tool allows the author to add new concepts to the ontology, define new relations and attributes, and all the features of each one. In addition, it also allows to define instance relations and instance attributes associated to the concepts in the ontology. Further development include a database consistency control as the previous tool. Figure 11 shows a snapshot of one screen of the interface.

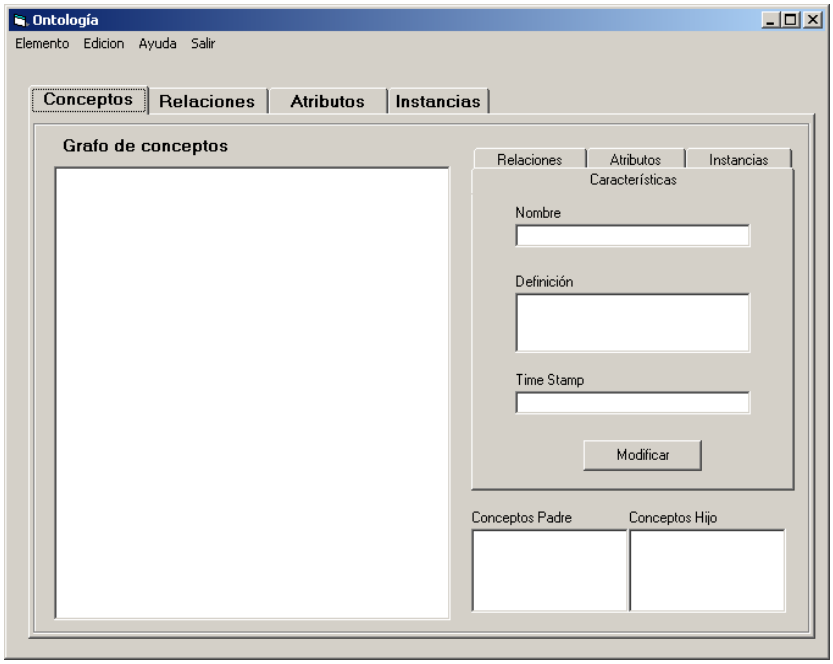

Figure 11. A Screen from the Ontology Tool Interface

The lexicon tool allows the author to add new terms as well as their features. It is in an early development stage and currently it is merged with the ontology tool.

Finally, the migration tool provides a way to interface the terminological database with the ontology and the lexicon. The migration is done with the supervision of an expert in the linguistic field selected. First, categories are migrated as concepts in the ontology, and the user is requested to map the category with an existing concept or a new one with the help of the existing concept graph. In addition, since categories represents relations between concepts, new instance relations are created for the meanings in categories. Terms from the terminological database are mapped to terms in the lexicon.

7. Conclusions and Future Work

We are in an very advanced step on the way to reach a sound and complete methodology to develop software systems for managing static linguistic knowledge bases. Based on this methodology we have built software tools for building and querying different kind of linguistic resources. Using these tools, information can migrate from one resource to another, thus permitting an easy integration among different knowledge bases. Naturally we must continue this work line taking into account more interesting conceptual and linguistic knowledge, augmenting the corresponding ontologies according to the adequate entity-relationship model, and adding the coherent target lexicons. The applications currently made embedding the conventional linguistic knowledge

bases will take advantage using these stronger integrated ones, and applications will come to new domains. The domains of NLP applications will wide. Besides managing these tools for languages learning is very promising, the application to education is a way to explore in the next future.

References

1. Casares, Ideological Spanish Dictionary.
2. C.J. Fillmore, and B.T. Atkins, “Toward a frame-based lexicon: The semantics of RISK and its neighbors”, Lehrer and Kittay, pp. 75-102, 1992.
3. B. Katzenberg and P. Piela, "Work Language Analysis and the Naming Problem", Communications of the ACM, Vol. 36, No. 4, June 1993.
4. D.B. Lenat, and R.V. Guha, “Building Large Knowledge-Based Systems”, Reading, Massachussets, Addison-Wesley, 1990.
5. María Moliner, Derivative Spanish Dictionary.
6. MikroKosmos, http://crl.nmsu.edu/Research/Projects/mikro/index.html
7. G. Miller, “WordNet: A Lexical Data Base for English”, Communications of the ACM, Vol. 38, 11, 1995.
8. A. Moreno, and C. Pérez, “Reusing the Mikrokosmos Ontology for Concept-based Multilingual Terminology Databases”, Proceedings of LREC2000, 2002.
9. S. Nirenburg, V. Raskin, and B. Onyshkevich, “Apologiae Ontologiae”, Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation, Center for Computational Linguistics, Catholic University, Leuven, Belgium, pp. 106-114, 1995.
10. R.S. Pressman, "Software Engineering. A Practitioner's Approach", McGraw-Hill, 1997.
11. A. Silberschatz, H.F. Korth, S. Sudarshan, "Data Base System Concepts", WCB/McGraw-Hill, 1996.
12. Y.A. Wilks, D.C. Fass, C.M. Guo, J.E. McDonald, T. Plate, and B.M.Slator, ”Providing machine tractable dictionary tools”. Machine Translation, 5, 1990, pp. 99-151.