Developing Dictionary Databases as Lexical Data Bases[1]

F. Sáenz and A. Vaquero

Departamento de Sistemas Informáticos y Programación,
Facultad de Informática, Universidad Complutense de Madrid,
E-28040 Madrid, Spain
{fernan, vaquero}@sip.ucm.es

Abstract

We propose to apply classical development methodologies to the design and implementation of Lexical Databases (LDB), which embody conceptual and linguistic knowledge. We represent the conceptual knowledge as an ontology, and the linguistic knowledge, which depends on each language, in lexicons. Our approach is based on a single language-independent ontology. Besides, we study some conceptual and linguistic requirements; in particular, meaning classifications in the ontology, focusing on taxonomies. We have followed a classical software development methodology for implementing lexical information systems in order to reach robust, maintainable, and integrateable relational databases (RDB) for storing the conceptual and linguistic knowledge. The result is a methodology to develop information systems for building and querying LDB (SV 02). Based on this methodology, we have developed software tools for authoring and consulting different kinds of linguistic resources: monolingual, bilingual and multilingual dictionaries. Conventionally, dictionaries are conceived for human use and lexical databases are conceived for natural language processing (NLP) applications. Our methodology leads to friendly usable dictionaries, but structurally prepared to be easily embedded in computer applications, as we show along the paper.

1 Introduction

Due to the immaturity of the knowledge representation topic, lack of standardization is broadly felt as a very undesirable state into the community around language resources (LREC 02). For instance, standard terminology for a common reference ontology is yet a goal to be reached. No doubt about what lexicon means, but ontology is differently understood in the computational linguistic literature. For instance, WordNet is mentioned as an ontology (USC 96), CYC is provided with a formal ontology (PRI 01), etc. Here, ontology, in a LDB, is the set of concepts in

[1] This work has been partially supported by the Spanish CICYT project number DPI2002-02924.

the domain of the base and the relationships that hold among them, without including linguistic knowledge, and common to all of the languages supported in the base.

Weak attention has been paid on topics about development methodologies for building the software systems which manage LDBs, and dictionaries in particular. We claim that the software engineering methodology subject is necessary in order to develop, reuse and integrate the diverse available linguistic information resources. Really, a more or less automated incorporation of different lexical databases into a common information system, perhaps distributed, requires compatible software architectures and sound data management from the different databases to be integrated. The database subject have already done a long way reaching a strong standardization, and supplying models and methods suitable to develop robust information systems. We apply RDB design methodologies to develop LDB consisting of ontologies and lexicons. The conceptual knowledge is represented as an ontology, and the linguistic knowledge, depending on each language, is stored in its lexicon.

Subjects about electronic dictionaries for diverse natural language processing applications have been extensively studied (ZOC 03), (WIL 90), (WIL96), as well as LDB (MIL 95), world knowledge bases (LEN 90), ontologies in general (ONT), ontologies for computational linguistics (NIR), and the like. But there are no references on how these information systems have been developed and upgraded along their life. Moreover, tools for managing ontology-based linguistic information systems have been described (MOR 02), but there is no a declared software engineering approach for the development of these tools.

We follow the classical RDB design based on the conceptual, logical, and physical models for building LDB, and software engineering techniques based on UML for building LDB interfaces (which are not described in this paper).

2 Conceptual and Linguistic Requirements

Conceptual and linguistic knowledge incorporated in computing systems devoted to NLP are relevant in the definition of a conceptual model for LDBs. Regardless of the language, the knowledge in the discourse universe is conventionally divided in two classes: conceptual and linguistic. Terms and sentences refer to concepts, but they have particular structural and morphological features in each language. All of this information is not available in any dictionary, electronic or not, although it is the objective in the most exigent ontology-based linguistic Knowledge Bases, such as MikroKosmos (MIK). (SV 02) provides more linguistic requirements we are interested in.

2.1 Lexical Databases

For a given language, we have a set of terms, meanings and categories holding certain relationships among them. Conventional LDB, such as WordNet (MIL 95), have term classification through synonymy (grouped in the so-called synsets). LDBs based on ontological semantics go beyond by playing the role of meaning taxonomy and supporting more complex semantic relationships (NIR 95). All of the relationships (meronymy, holonymy, hypernymy, hyponymy, and so on) represented in the more complete lexical databases, such as WordNet or EuroWordNet (EWN), are also represented in ontology-based databases, such as MikroKosmos; but in this case, all of the concepts and their relationships are present in the ontology, while each lexicon has the terms for each language and their linguistic arguments, as well as the links with the concepts into the ontology. The mapping between ontology and lexicon is the key for successfully coordinate all of the lexical and semantic relationships. This approach does full separation between ontology and lexicon. If we now think of several languages, the same ontology applies for each one of the lexicons.

Any other approaches has been adopted. Each one of them leads to a more or less complex LDB structure. We claim for the approach ontology-lexicons as the most appropriated to reach a simple, robust and controlled LDB structure, prepared to be reused in different applications and integrated with another ones with the same structure.

The architecture ontology-lexicons is criticized in (POL 03), given that each language has its own lexical semantics. Then, strictly speaking, there is no one single ontology independent of the considered languages. In favor of our position, we argument that the fact of the nonexistence of one single ontology common to diverse languages is independent of assuming one imposed undesirable a priori hierarchy, which is considered in (POL 03) as unavoidable considering the common ontology approach. But in our methodology, the hierarchy (taxonomy) is incrementally created when building the LDB. For a monolingual database (French in the case of the DiCo LDB), there is only one ontology; thus, there is no problem. However, certain problems could arise in multilingual LDB, because the boundary between ontology and lexicon does not appear clearly always. There are many ways to face up these problems considering other approaches different from ours, when the ontological semantics is distributed among the different languages at multiple levels. For instance, in the Papillon project (MAN 03), the different languages are linked to a common dictionary of meanings (axies in French). In the EuroWordnet project, the different WordNets (one for each considered language) are linked by two levels of common concepts, and the resulting structure is not appropriated for the multilingual applications. In MILE (ABB 02), SIMPLE templates play the role of ontologies; so the resulting LDB structure is more complex than that resulting from the approach ontology-lexicons.

We adhere to the criterium from (MAH 95) conceiving ontology as a language-neutral body of concepts. In this case, the problems can be solved putting in each specific lexicon the own lexical-semantic information required, which is not present in the common ontology (VIE 98); so the ontology is the conceptual model of the domain and each lexicon is linked to the same ontology. From this approach, the system design to develop LDB is enhanced in robustness, because an architecture with two abstraction levels is reached.

From this approach we apply very carefully the RDB techniques to reach a methodology assuring a sound and simple structure of the LDB, and a controlled way for building any particular LDB through an administration interface. This work is indeed previous to the formal definition of an interlingua (FAR 04). We are far from reaching this goal, but there are a lot of NLP applications, not only monolingual ones, that do not need formally and completely represent the text meaning. We claim for reaching an interlingua in the future from LDB conceived from the ontology-lexicons approach and developed with our methodology.

Another central idea in this work is to develop for each group of applications one LDB, the most appropriated one. Certain applications are more exigent of linguistic resources than other ones. Why to use the same LDB for no matter what application?. This vision contemplates, besides our methodology to build different LDB, building subsets of LDB already build as 'views' of the DB; in this case the LDB has to have been developed from the ontology-lexicons approach. We claim for this way in order to integrate different LDB.

2.2 Our LDB for Dictionaries

In this approach, relationships among terms from different languages come from considering jointly the involved Ontology-Lexicon schemes, as we will see later when considering the bilingual dictionary. In the dictionary here considered, the ontology only consists of one relationship which gives tree-structure to the conceptual taxonomy. A taxonomy is a natural structure for meaning classification. Each node in the taxonomy corresponds to a category. In principle, every category in the taxonomy can have meanings, regardless of its taxonomy level. It must be noted that every category in the taxonomy 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 belonging to several predefined ones should be avoided, in order to reach a compact relationship as the taxonomy structuring backbone. We have developed dictionaries without overlapped classifications (RK 02), and only permitting tree-structured taxonomies. Since a meaning can belong to different categories, the extensional definition of categories is hold (SV 02).

When consulting or building dictionaries, there are a number of advantages in classifying meanings as taxonomies. First of all, meaning taxonomy is a useful

facility for an electronic dictionary, because meaning classification embodies additional semantics, which provides more information to the user than usually provided. As long as we know, this kind of facilities (meaning classification), normally used in conceptual modeling through ontologies (MCG 00), has not been implemented before into dictionaries.

3 Conceptual Model of the Terminological Database (TDB)

There are different TDBs built for different purposes. Some of them have incorporated the ontology structure, and so, they could possibly be used for the pedagogical goals proposed above. But there are a lot of difficulties when intending to do this, not being the less the fact that these very large databases are yet complete or almost complete. So only the tools for building terminological databases are needed. Moreover, the development of this kind of tools must be made taking into account the pedagogical goals which have not been the case of the LDB already built.

Our work in developing the tools is based on a sound conceptual model for the terminological database 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'll use unique representations for them.

The entity-relationship model is used to describe the conceptual model we propose, shown in figure 1. In this figure, 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). Undirected lines also connect attributes to entity sets. Relationship set and entity set names label each diamond and box, respectively.

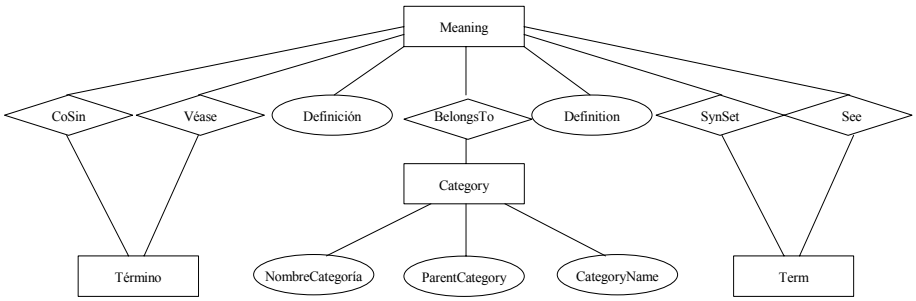

Figure 1. Entity-Relationship Model for an English-Spanish TDB

For the sake of clarity and conciseness, in this figure we show an instance of a multilingual terminological database for only Spanish and English languages, although we have extended for multilingual support (SV 02). 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 relationship set SynSet denotes the English synonym set. The entity set Term represents all the English terms that compose the terminological database. The relationship set between Meaning and Term 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 polysemic. The entity set Category denotes the category each meaning belongs to. 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 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. The remaining entity and relationship sets (CoSin, Véase, Término) are homologous to the ones in the other language (SynSet, See, Term.)

The logical and physical models for the development of any terminological database following the principles above expressed have to be based on this conceptual model.

4 Functionalities of the Tools

4.1 The User Tool

We have developed a user tool, a query interface which allows us to easily recover the information about both English and Spanish terms as well as their rela-

tionships from the so-called 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, …) as shown in (SV 02).

The Start window of this tool allows the user to select the base language (i.e., the source language for translations and for representing dialogues) among the available languages by pressing its button (from now on, we consider a bilingual dictionary so that it is unnecessary to select the source language or the target language.)

This action pops up the Semantic Category window; its left pane shows the semantic categories structured as a tree, and the right pane, all the words under the highlighted semantic category. The total number of terms is showed on top of the right pane. The nodes in the tree can be clicked in order to expand or contract semantic categories subtrees. A text box is used for term lookups so that the closest word to the substring typed is shown in the right pane. Pressing Enter or double-clicking the highlighted word yields to the Query window. This window shows the relevant information about the selected term: its definition, comments, the list of semantic categories it belongs to (the one corresponding to the shown definition is highlighted), the synonym set and the list of related terms. It also displays a navigation history. It is possible to select another semantic category in this window, which results in updating all the relevant information. Direct access to the terms in both the synonym and related terms windows is allowed by double-clicking.

The Semantic Category window has a control box with buttons to activate the return to the Start window, navigate backwards, translate the selected word, print, and exit the interface. The Translate button offers one of the main functionalities of this interface, i.e., the translation from the (source) base language to the target language and, when pushed, it pops up the Translation window. This window shows a first field for the term in the first language, and a second field for the term in the second language. There are also navigation buttons for searching other terms in the same semantic category under an alphabetical order. It is possible to translate from the first or from the second language by using two buttons which express the two possible translation directions. Also, the Go to buttons allow us to go to the Semantic Category window for the selected term. This completes the overall description of the functionalities of the user tool.

4.2 The Author Tool

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), and it consists mainly of one Author window. It has several management areas which are explained next.

Semantic Category Management Area This area is intended for managing all the operations related to semantic categories. It has several controls: a hierarchical view of the semantic categories (with expand/collapse functionality), text fields for the semantic category names (English and Spanish), and the buttons Add Category, Delete Category, and Modify Category. The insertion point when adding a new semantic category is the highlighted semantic category, and the Spanish and English texts for the semantic category name must be typed in the aforementioned text fields.

Meaning Management Area The area for meaning management consists of two lists for the meanings in both languages and the buttons Add, Delete, and Modify for addition, deletion, and modification of meanings, as well as buttons for edition (Copy and Paste buttons.) These lists shows the meanings in the form Term -> Definition for the highlighted category, so that one can see several meanings for the same term. Moreover, when a pair Term -> Definition is selected, the corresponding Term -> Definition translation is automatically highlighted; there is a one-to-one mapping between meaning representation in all the languages. It should also be noted that meanings, which are language independent, are shown with the *best* representation we have in a given language, i.e., a pair Term -> Definition, since there are no other pair Term -> Definition2 with the same meaning (note that is the same term in both pairs.)

Synonyms and Related Terms Management Area This area has four lists for the synonyms, and related terms in both languages which correspond to the highlighted meaning in the Meaning Management area.

Database Control Area This area contains a button which is used to obtain a report about consistency of the database. Consistency detection reports about lack of textual definitions for terms, and other inconsistencies (circular references) and omissions (lack of related terms via relationships See and SynSet). This is quite important when authoring dictionaries, since a dictionary cannot be consistently built at each step, but it is constructively built from terms to relationships between terms (polysemy, synonymy.)

5. Conclusions

Continuing with the refinement of our development methodology of information systems for lexical databases, we have followed an elaborated and well sound design method. The design is based on the ontological semantics approach, and we have signaled the advantages of this approach in face of the non-ontological one. The design has been tested and used to complete the development of certain infor-

mation systems to build and consult monolingual, bilingual and multilingual dictionaries.

Of course, the advantages of applying software engineering principles and methods to information systems for lexical databases are evident. Moreover, by using the resulting tools, the LDB authoring is a friendly simple task, and the inserted information has to accomplish certain constraints (consistency, non recurrence, ...) controlled by the system, helping the authoring process (avoiding violation of hard constraints and reporting the violation of soft constraints). Besides, the integration of diverse LDB built with these tools is assured by the migration tools developed for this purpose. In addition, the resulting dictionaries are friendly usable and supply very useful semantic information to the reader.

References

(ABB 02) Atkins S., Bel N., Bertagna F., Bouillon P., et al (2002) "From Resources to Applications. Designing TheMultilingual ISLE Lexical Entry". In Proceedings of LREC 2002, Las Palmas, Canary Islands, Spain.

(EWN) http://www.uva.nl/EuroWordNet.html

(FAR 04) D. Farwell "Intermediate Representation". Seventh Interlingua Workshop AMTA'04: Determining Interlingua Utility for Machine translation. Washington, DC, October, 2004.

(LEN 90) D.B. Lenat, and R.V. Guha, "Building Large Knowledge-Based Systems", Reading, Massachussets, Addison-Wesley, 1990.

(LREC 02) Workshop on "International Standards of Terminology and Language Resources Management", Las Palmas de Gran Canaria, June, 2002.

(MAH 95) K. Mahesh, and S. Nirenburg, "A situated ontology for practical NLP". IJCAI'95. Montreal, August 19-21.

(MAN 03) M. Mangeot-Lerebours, G. Sérasset, M. Lafourcade. "Construction collaborative d'une base lexicale multilingue. Le projet Papillon". TAL, Vol. 44 – 2. 2003

(MCG 00) Deborah L. McGuinness. "Conceptual Modeling for Distributed Ontology Environments", Proc. of The 8th Int. Conf. on Conceptual Structures Logical, Linguistic, and Computational Issues (ICCS 2000), Darmstadt, Germany, August 14-18, 2000.

(MIK) MikroKosmos, http://crl.nmsu.edu/Research/Projects/mikro/index.html

(MIL 95) G. Miller, "WordNet: A Lexical Data Base for English", Communications of the ACM, Vol. 38, 11, 1995.

(MOR 02) A. Moreno, and C. Pérez, "Reusing the Mikrokosmos Ontology for Concept-based Multilingual Terminology Databases", Proc. of LREC, 2002.

(NIR 95) S. Nirenburg, V. Raskin, and B. Onyshkevich, "Apologiae Ontologiae", Proceedings of the Sixth International Conference on Theoretical and Methodo-

logical Issues in Machine Translation, Center for Computational Linguistics, Catholic University, Leuven, Belgium, pp. 106-114, 1995.

(NIR) S.Nirenburg and V.Raskin, "Ontological Semantics", In http://crl.nmsu.edu/Staff.pages/Technical/sergei/book.html

(ONT) http://www.ontology.org/main/papers/iccs-dlm.html

(POL 03) A. Polguère, "Etiquetage sémantique des lexies dans la base de données DiCo". TAL, Vol. 4 – 2. 2003.

(PRI 01) U. Priss, "Ontologies and Context". Midwest Artificial Intelligence And Cognitive Science Conference. Oxford, OH, USA, 2001

(RK 02) C. Raguenaud and J. Kennedy, "Multiple Overlapping Classifications: Issues and Solutions". 14th International Conference on Scientific and Statistical Database Management (SSDBM'02). Edingburgh, Scotland, 2002.

(SV 02) Sáenz, F. & Vaquero, A. "Towards a Development Methodology for managing Linguistic Knowledge Bases". Proceedings ES'2002. Springer-Verlag, 2002. pp 453 – 466.

(USC 96) M. Uschold and M. Gruninger, "Ontologies: principles, methods, and applications". Knowledge Engineering Review, Vol. 11, 2. 1996, pp 93-155.

(VIE 98) E. Viegas, "Multilingual Computational Semantic Lexicons in Action: The WYSINNWYG Approach to NLP". Int. Conference on Computational Linguistics, ACL. Montreal, 1998.

(WIL 90) 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.

(WIL 96) Y. Wilks, B. M. Slator, and L.M. Guthrie, "Electric words: Dictionaries, Computers and Meanings". MIT Press. Cambridge, 1996.

(ZOC 03) M. Zock and J. Carroll "Les dictionaires électroniques". TAL, Vol. 44, 2. 2003.