A Human-Learning Environment for Building and Querying Electronic Dictionaries

A. Vaquero, F. Sáenz
Departamento de Sistemas Informáticos y Programación, Universidad Complutense de Madrid
E-28040 Madrid, Spain {vaquero, fernan@sip.ucm.es}

Key words: Computer Based Human-Learning, Computer-Human Interfaces, Constructivism, Language Mastery, Electronic Dictionaries, Lexical Database, Ontology

Abstract: This paper presents the current state of software tools we have developed for improving the language mastery of students and people, in general. These tools involve linguistic concepts needed in every subject and in a broad range of education levels. Lexicon concepts useful for language learning are involved in dictionaries and other kinds of linguistic resources. These concepts (vocabulary, meanings, semantic categories, semantic relationships, and taxonomy) are pointed out. Claim is made for using new environments and computer-human interfaces based on these concepts which define the pedagogical goals. Envisioning the foreseen collaborative and individual instructional tasks, claim is made for the constructivist model of learning. Authoring and consulting electronic linguistic resources are the main tasks to reach the defined goals. The conceptual model appropriate for developing any software system taking into account all these principles is discussed, and the reached entity-relationship model is presented. We further present the developed tools for creating bilingual linguistic resources, which could allow to reach the foreseen learning goals. More sofisticated tools for advanced users interested in linguistics are been developed.

1. Introduction

Lack of standardisation is broadly felt as a very undesirable state into the community around ontologies, lexicons, and so on. Attention has not yet been paid on 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 necessary.

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 [FIL 92].

Subjects about electronic dictionaries for diverse natural language processing applications have been extensively studied [WIL 90], as well as Lexical Databases [MIL 95], World Knowledge Bases [LEN 90], ontologies [MIK 02], 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 [MOR 00], but there is no a software engineering approach for their development. To develop our tools 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. Linguistic concepts embodied in the lexical and ontological resources are firstly exposed in section 2, because of their relevance in building different Linguistic Data Bases, such as Electronic Dictionaries. The next sections briefly expose the different conceptual models we present for several linguistic resources. For all of them, we have followed the conventional 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 de-

velops a conceptual model for an ontology (we have selected MikroKosmos [MIK 02]). 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. In section 7 we define a methodology to build collaboratively complex dictionaries, and it is said how to apply it to build 'the' Spanish–English dictionary of Information and Communication Technologies. In section 8 we face the problem of the mother tongue mastery and languages learning and the contribution to solve it by applying our tools. Finally, Section 9 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 [MAR 02] 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 [CAS 02].) Semantic categories not also allow meaning classification, but the more meaningful taxonomy of meanings. Conventional lexical databases, such as WordNet [MIL 95], have term classification such as synonymy (grouped in the so called synsets.) Ontologies go beyond by playing the role of meaning taxonomy [NIR 95]. 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. 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). Therefore, semantic categories provide classification for meanings, and such classification can be arranged in a taxonomy. It is commonly acknowledged that the best order for lookups is lexicographic. 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 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.

2.2. *Polysemy and Synonymy*

In every language there exists the well known naming problem [KAT 93], 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.

2.3. *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.. Then, we have one more meaning in each category. This meaning is the intensional definition of the category. For a given language, we have a set of terms that holds the relationships with categories and meanings. 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 .

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. 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 Mikro-Kosmos 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. Design Models 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[PRE 97, SIL 02]).

We have also developed the logical and physical models for the development of our terminological databases, 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. We developped also the logical and physical models of the tools from the entity-relationship model above mentioned.

Logical Design of the Spanish-English Dictionary Management Tools

In addition to the entity-relationship diagram, a result of this stage is the definition of some constraints. Mapping cardinalities are taken into account, and also the participation of entities in relationships with elaborated participation cardinality. For instance, a definition and a synonymous set must be defined for each meaning, and also a term have to be related because when one adds a new meaning both a term and a textual definition may define it in each of both languages.

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

We have developed a conceptual model the terminological database for a multilingual dictionary, where Language is an Entity . Figure 1 shows the entity-relationship model.

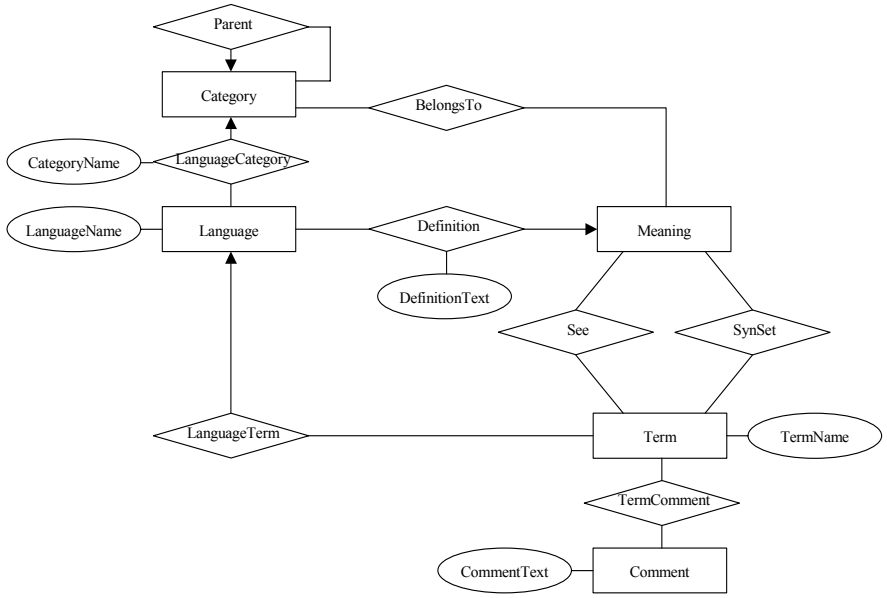

Figure 1. ***Entity-Relationship Model for a multilingual TDB***

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.

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.

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 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).

Note that definitions are 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 [MIK 02] 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 [MIK 02] can be viewed as a directed graph with concepts as nodes. There are semantic relationships among nodes. 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. 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 2 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.

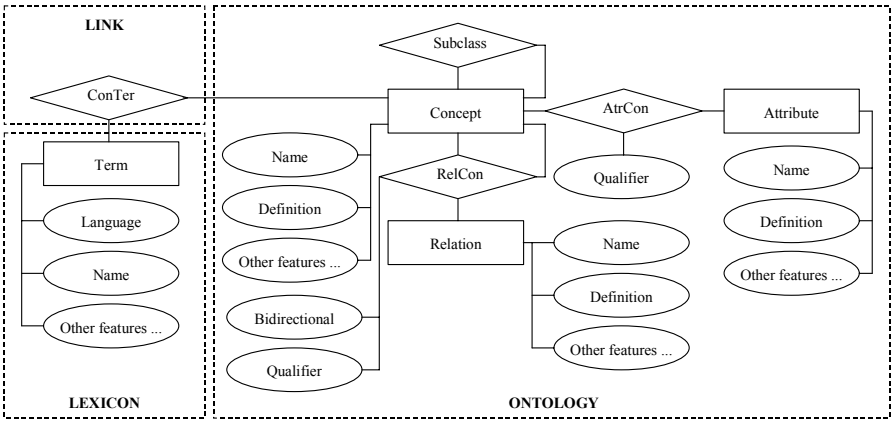

Figure 2. ***Entity-Relationship Model for the MikroKosmos Ontology***

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, …).

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.

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.

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. Methodology to build big dictionaries

The methodology for building a terminological data base system consists of three main tasks: First, gathering information; its goal is to provide all the needed information for the electronic dictionary, i.e., terms, term translations, and term definitions. Second, building the terminological data base in order to be able to support the gathered information. Finally, developing an application interface for the end user.

Up to now, we have worked mainly in the first task for an electronic Spanish-English dictionary of Information and Communication Technologies project. Its activities are the following: looking for information sources, gathering interesting information, and providing correct translations and definitions under our criteria. The first activity is organised by the team director, who assigns different search categories for each team member. Therefore, each member has to look for information sources all over both the on line and printed worlds in its search category (different glossaries, dictionaries, forums, ... related to the Informatics world).

The second activity, gathering interesting information, relies on the responsibility of each team member. We will accomplish the last activity, providing correct translations and definitions, in two pipelined stages.

8. Learning Language Mastery

When weak domains in the student skills and knowledge are detected, it is mandatory to fill the gap by applying appropriate computer-based environments. A specially weak domain is Language. There exist a worrying lack in the mother tongue mastery of the young and not so young people all around the world. The key part of the language misunderstanding is lexicon. There is experimental evidence of reading comprehension dependency of the vocabulary [JOH 78, THO 73].

In order to improve the level of language mastery, every pupil ought to handle specific tools with facilities for creation, consulting and modification of language parts. Firstly allowing the student more easily and quickly look up terms. Moreover, the computer could allow the development of new tasks with clear pedagogical goals according to the learning model based on constructivism. This implies a new approach to the concept of electronic linguistic resources other than the electronic counterparts of the printed ones [WIL 90] [WIL 96].

The global pedagogical goal to be reached is word meaning [QUI 67]. The relations between two words of different languages are given through the ontology and its connection to both lexicons. All these goals can be reached following a constructive and collaborative way among students and the teacher in the classroom. This could only be efficiently done with appropriate tools and friendly usable interfaces as a whole responsive environment [ZEL 97].

In order to situate our tools in their correct instructional place, one must distinguish between constructivist learning in user controlled environments (fully free environments) and navigation in hypermedia ones [NOR 94].Our tools belong to the first one of these two models of learning, the second one being more appropriate for learning other parts than lexicon [GOL 96]. Nonetheless, both are complementary and not absolutely separate [TEU 96]. As the second one has been extensively treated in the past [FER 99], the first one is highlighted here.

Our tools fill a niche for improving the language mastery of students in every subject, and in a broad range of education levels. In addition to the above need, a pedagogical goal is to reach a customisable personal dictionary that allows other functionalities such as personal notes, taxonomies, etc., which are not included in conventional dictionaries. Besides the advantages noted above in using and authoring electronic dictionaries with ontologies for language learning, we also consider foreign language learning under, first, the same lexicon meaning point of view, and second, the relationships between the foreign language and the mother tongue, that is, a multilingual electronic dictionary with support for ontologies. So the student can comprehend the language independence notion of meanings, so that semantic categories can independently be defined from the language.

9. 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

[CAS 02] Casares, Ideological Spanish Dictionary.

[ERI 94] F.J. Erickson, and J.A. Yonk, ”Computer Essentials in Education. The teaching tools”. McGraw-Hill Book Co., 1994.

[FER 99] A. Fernández-Valmayor, C. López-Alonso, S. Arlette, and B. Fernández-Manjón, “The Design of a Flexible Hypermedia System: Integrating an Interactive Learning Paradigm for Foreign Language Text Comprehension”, International Working Conference on Building Electronic Educational Environments, IFIP, Irvine, California, pp. 51-65, 1999.

[FIL 92] 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.

[GOL 96] S.R. Goldman, ”Reading,Writing,and Learning in Hipermedia Environments”, Cognitive Aspects of Electronic Text Processing (Ed.H.Van Oostendorp and S. Mul), Norwood, NJ. Ablex Publications, 1996.

[JOH 78] D.D. Johnson, and P.D. Pearson, ”Teaching Reading Vocabulary”, Ed. Holt, Reinhard & Winston, New York, 1978.

[KAT 93] B. Katzenberg and P. Piela, "Work Language Analysis and the Naming Problem", Communications of the ACM, Vol. 36, No. 4, June 1993.

[LEN 90] D.B. Lenat, and R.V. Guha, “Building Large Knowledge-Based Systems”, Reading, Massachussets, Addison-Wesley, 1990.

[MAR 02] María Moliner, Derivative Spanish Dictionary.

[MIK 02] 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”, Proceedings of LREC2000, 2002.

[NIR 95] 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.

[NOR 94] K. Norman, “Navigating the educational space with HyperCourseware”. Hypermedia, Vol. 6, enero 1994.

[PRE 97] R.S. Pressman, "Software Engineering. A Practitioner's Approach", McGraw-Hill, 1997.

[QUI 67] M. R. Quillian, “Word Concepts: A Theory and Simulation of Some Basic Semantic Capabilities”. Brachmen, R. J. y Levesque, H. J., Eds., Reading in Knowledge Representation. Morgan Kaufman, 1.967.

[SIL 02] A. Silberschatz, H.F. Korth, S. Sudarshan, "Data Base System Concepts", WCB/McGraw-Hill, 2002.

[TEU 96] P. Teusch, T. Chanier, Y. Chevalier, D. Perrin, F. Mangenot, J.P. Narcy, and J.de Saint Ferjeux, ”Environnements interactives pour l’apprentissage en langue étrangère”. Hipermedias et Apprentissage, 3 (Ed. E. Brouillard), 1996, pp. 247-256.

[THO 73] R.L. Thorndike, ”Reading Comprehension Education in Fifteen Countries”, Ed. Wiley, 1973.

[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 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: Diccionaries, Computers and Meanings". MIT Press. Cambridge, 1996.

[ZEL 97] D. Zeltzen and R. K. Addison, “Responsive virtual environments”, of the ACM, Vol. 40. N. 8, August, 1997.