LEARNING LINGUISTIC CONCEPTS THROUGH THE CONSTRUCTION OF DICTIONARIES WITH A DIRECTED ACYCLIC GRAPH SHAPED TAXONOMY

A. Vaquero, F. J. Alvarez, and F. Sáenz
Departamento de Sistemas Informáticos y Programación, Universidad Complutense de Madrid, Facultad de Informática, José García Santesteban s/n, E-28040 Madrid, Spain. Tel.+34 91 394 7622, Fax:+34 91 394 7529. E-mail: vaquero@sip.ucm.es, francisco_alvarez@fdi.ucm.es, fernan@sip.ucm.es

Abstract: In general the global pedagogical goal of the creation and use of dictionaries is the mastery of languages. The specific goal is to facilitate the assimilation of basic linguistic concepts: lexicon, sense (meaning), semantic categories, lexico-semantic relations, etc. In this paper, we describe the linguistic concepts that constitute the pedagogical goals to attain, as well as the constructivist model of learning. The conceptual knowledge linked to the terms is arranged in taxonomies. However, in our previous work, the concepts and their relationships were represented as simple hierarchical trees. Nevertheless, these taxonomies are quite limited and thus do not capture the complexities of reality in a given area, as they only allow a category to have one parent category. We thus opted for a directed acyclic graph (DAG) to represent the conceptual knowledge as this structure allows a more specific category to be the child of multiple broader categories. With these requirements we have developed tools to direct the activities that allow building and querying monolingual dictionaries based on a DAG-shaped taxonomy. Using these tools to build a dictionary, users can learn in a simple and natural way some semantic aspects that are necessary for the mastery of languages.

Keywords: Linguistic Concepts, Dictionaries, DAG, Taxonomy, Ontology, E-R Model, Tools
1. INTRODUCTION

One the most important values in the classroom is language. The teaching and learning of any subject depends on the communication, direct or through technological means, between teacher and student. But language itself is a technically complicated subject to teach and master. When a weak domain of knowledge and abilities is detected among the student population, as language is, this deficiency must be dealt with the appropriate information technology environments. Conventional resources, such as dictionaries, glossaries, thesauri, encyclopaedias, etc., are not enough, as shown through the generalized fact of its scarce utilization. Information technology can help thus to, in particular, master the native language.

Having cleared the pedagogical goal of each activity, the use of computers in school must be controlled by the teacher (Cuban, 1987). However, the teacher must possess, besides a general culture about informatics a specific formation in electronic educational environments (Erickson and Yonk, 1994). However, when the teachers try to use computers in the classroom, several problems are detected, among which we can highlight: the quality and accessibility of the software, the selection of the appropriate software for classroom teaching, the scarce reliability of the revisions and evaluations of the educational software, the difficulty to integrate the existent software into the teaching process and the limited time available before the teacher can actually use information technology in the classroom (Hodgson, 1994). These problems become more serious when it comes to teach a language, the learning of which follows an imposed constructivist model (Cabrera, 1995), based on cognitive theories (Ausubel, 1968; Posner, 1989), that induces a new form of teaching (Karat, 1997; Tobin and Tippings, 1993). Consequently, new information technology based resources must be built in order to accomplish activities that can raise the motivation needed to understand and use the natural language in a correct way. According to constructivism, electronic dictionaries (Wilks et al., 1990) motivate more than its paper counterparts. To improve the level of language mastery, each student should use specific tools that would allow him to build, query and modify pieces of the language. The global pedagogical goal to attain is the sense (meaning) of words (Quillian, 1967), because the key to a better understanding of the discourse is the lexicon. There is experimental evidence of the dependency between reading comprehension and vocabulary (Thorndike, 1973; Johnson, 1978). The definition of the sense of a term constitutes the task through which it is possible to understand that the meaning of a word depends on the meaning of other words, just like with the classification of meanings in semantic categories. Specific pedagogical goals are the lexical relations polysemy and
synonymy and their implications in classification. All these objectives can be attained via a constructivist procedure and through the collaboration between students and teachers in the classroom. This mission must be accomplished with appropriate reactive tools (Zeltzen and Addison, 1997) provided with user friendly interfaces.

Semantic categories are organized and indexed based on a classification principle that describes a hierarchy of concepts: a conceptual taxonomy. In the natural language processing (NLP) area, these conceptual taxonomies become semantic taxonomies composed of semantic categories. These taxonomies, although they don’t fully cover the knowledge representation problem, provide a solid base for NLP (Jacobs, 1991). This approach has been followed for the construction of tools for the creation and management of dictionaries (Vaquero, Sáenz and Lopez, 2003).

However, the preceding tools only allow to create tree shaped taxonomies. This represents a limitation because complex classifications can be obtained where a category can have more than one parent. For example, a novel can be of genre fiction and non-fiction at the same time. The fact is that these classifications do exist and are seen to be useful and, therefore, cannot be ignored (Raguenaud and Kennedy, 2002).

In this article, we propose software tools to create, manage and query monolingual dictionaries based in a DAG-shaped taxonomy, as a means for continual learning and improving the student’s knowledge of a language until a high degree of mastery has been acquired. This taxonomy is based on a single relationship (e.g., IS A) that is implicitly represented and, even though there can be multiple inheritance, it is an implicit one, as opposite to the one we have when frames are used to represent the taxonomy, as in (Mahesh and Nirenburg, 1995). The relationship is also implicitly stated by the author and he should therefore create the dictionary according to it. However, he has no control through the tools for ensuring the semantic correctness of the taxonomy with respect to the intended relationship. It should also be noted that our single (directed) relationship forms a partial order (a reflexive, antisymmetric, and transitive relation). This condition is particularly useful for learning as direction facilitates understanding (Raguenaud, 2000). These tools have been developed following the methodology described in (Sáenz and Vaquero, 2002, 2005a, 2005c). We also introduce the conceptual model from which the tools were developed.

However, to situate our tools in their right place, we must distinguish between constructivist learning in user controlled environments and the multimedia navigational ones (Norman, 1994). Our tools belong to the first of these learning models. The second is more appropriate for learning other parts of the lexicon (Goldman, 1996), however, both types of tools are
complementary and they are not separable in any way (Teusch et al., 1996; Fernandez-Valmayor et al., 1999).

The rest of the article is organized as follows. In section 2, the linguistic concepts considered as learning goals are pointed out. In section 3, the structure of the definitions when using taxonomies is established. In section 4, the importance of the meaning classification in taxonomies is highlighted. In section 5, the conceptual model for monolingual dictionaries is presented. In section 6, the novel aspect of the latest developed tools and their interfaces are explained. In section 7, some conclusions, as well as some future work, are outlined.

2. LINGUISTIC CONCEPTS INCORPORATED AS PEDAGOGICAL GOALS AND HOW TO REACH THEM

In this section, the diverse concepts that are represented in the conceptual models of the previously developed dictionaries are pointed out. In all the models, a series of linguistic concepts are represented, these concepts are important for NLP purposes, as well as for human comprehension. We are only going to mention them here as they already have been explained before (Vaquero, Sáenz and Lopez, 2003; Sáenz and Vaquero, 2005a, 2005c). Such concepts are: order, classification, ontology, lexical relations (polysemy and synonymy) and a single semantic relation among concepts.

How to reach this pedagogical goal is conceived through the construction of dictionaries using the tools developed, particularly the learning tools that are shown in section 6.

In general, lexical databases (LDB) are used as linguistic resources with two purposes: for human consulting and as embedded static information in NLP applications. Each instance of a LDB is built by lexicographers and other linguistic experts. We believe that this constructivist effort implies the learning of the improved concepts and, thus we propose that the learner builds dictionaries about the domains to master using these tools (Vaquero, Sáenz and Barco, 2001). Moreover, a team of authors can build a specialized dictionary, dividing the work by categories that are at the same hierarchic level, therefore promoting collaborative work between students (Vaquero, Sáenz and Barco, 2000).
3. STRUCTURE OF THE DEFINITIONS WHEN USING A TAXONOMY

Even though language is still the preferred vehicle to structure definitions, there is the tendency to complement or substitute the definition by representations that allow to represent the information in a more exhaustive way. This reflects the interest in searching other graphical and semiotic forms of representations that go beyond the conventional text that follows every one of the entries in a dictionary and underlines the conceptual importance of the definition, or even its absence. The use of a concept system could be a sufficient condition to comprehend, which means that, sometimes, many terms are clearly understood and do not need a definition and, sometimes, many terms are easily identified through the definitions they establish with other terms, and thus they do not need any kind of identification or additional explanation (García de Quesada, 2001).

The concepts require the support of definitions to enable terminology users and associated software applications to gain access to them. The definitions thus created serve to restrict the sort of entries that can be admitted as falling under the corresponding concepts (Smith, Ceusters and Temmerman, 2005).

The definition of a concept is very important, as it describes it in an exhaustive or partial way, and allows to indirectly delimit and fix the content of a concept through references of other concepts. It establishes a sense of equation between what must be defined (definiendum) and that by which something is defined (definiens).

Conceptual systems are based mainly on the extension and intension of concepts, this makes it possible to talk about extensional and intensional definitions.

According to Wüster (Wüster, 2003), before we can assign a term to a concept, we must first delineate it. To delineate means: to list the totality of "characteristics" which form a concept's content or intension, that is, the set of semantic features that define the class denoted by the category. The intensional definition describes a concept by its most essential, relevant and intrinsic characteristics. Each definition of the meaning (sense) of a term is intensional, because of that, it must be very precise.

3.1 Extensional Definition

The extensional definition of a concept is defined as the set of known subordinated concepts for which the intension of the concept is true. The more semantic features the intension of a concept has, the more limited its
extension is. More semantic features are needed to define “seagull” than to define “bird”, even though there are more birds in the universe.

Extensional definitions could be very useful for the comprehension of a concept to define. However, sometimes it is extremely difficult to enumerate all the examples of a concept. For example, the category “planet”, by its intensional definition, allows the incorporation of a new planet in its extension when it is discovered.

Because of this, the conceptual structure must contemplate at the same time the intensional and extensional definition in each semantic category.

3.2 Relationships among Categories, Meanings and Terms

Here we do some remarks about the relationships among 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 1 shows three categories (C₀, C₁ and C₂). In this figure, the set of meanings {M₁₁, M₁₂, M₃} is the extensional definition of category C₁ and the set of meanings {M₂₁, M₂₂, M₃} is the extensional definition of category C₂. As stated in section 3.1, each category has a meaning described by a definition (its intensional definition), thus, C₀, C₁ and C₂ are the meanings of categories, C₀, C₁ and C₂ respectively and at the same time, meanings C₁ and C₂ are the extensional definition of category C₀.

Now, each meaning has one or more terms associated. The term T₂ is associated to meanings M₁₁ and M₂₁, which respectively belong to categories C₁ and C₂. Terms T₄ and T₅ are assigned to meaning M₂₂ that belongs to category C₂. The term T₅₀ is associated to meaning C₀ and TC₁ and TC₂ to C₁ and C₂, respectively. Finally, there is the term T₃ that is assigned to M₃, which belongs to both C₁ and C₂ at the same time. The fact that we can have M₃ as part of both C₁ and C₂ represents the difference with respect to our previous work.

As for the lexical relations, polysemy is present in T₂ and synonymy in T₄ and T₅. T₁ is neither polysemic or synonym. Finally, as it can be deduced from figure 1, each category must, at least, have one associated term denoting it.
4. ADVANTAGES OF THE CLASSIFICATION OF MEANINGS IN TAXONOMIES

The taxonomy of meanings is a very useful feature for an electronic dictionary, because it provides the user making the query, information not only about the intensional definition of a concept, but also, information about the conceptual hierarchy. The explicit representation of concepts and their relationships facilitates graphical representation of knowledge structures, thus helping the learning process by providing the kind of “conceptual maps” advocated by numerous educational psychologists (Meyer et al., 1992).

As already said, another advantage is collaborative work among students, which our tools promote but do not apply a collaborative methodology; however, they can be upgraded to follow an approach as the one in (Vaquero, Sáenz and Barco, 1999). This methodology, intended for professionals, can be adapted to the learning environment.

From the educational point of view, the goal is not to develop a general dictionary, but specialized dictionaries restricted to one semantic or linguistic domain, in order to make the categorization of meanings and the definition of categories simpler.
5. CONCEPTUAL MODEL FOR MONOLINGUAL DICTIONARIES

The classical relational model has been proposed to represent dictionaries (Nakamura and Nagao, 1988; Tiedemann, 2002). However, none of these projects uses an appropriate conceptual data model as a point of departure nor do they follow a software engineering approach for the construction of the database and the tools for its creation and management. In addition, domain knowledge isn’t explicitly represented.

Thus, we propose a sound and simple conceptual model for monolingual dictionaries where domain knowledge is represented in an explicit and structured way.

The proposed model represents a refinement over previous models (Vaquero, Sáenz and Lopez, 2003; Sáenz and Vaquero, 2005a, 2005c) and it is the point of departure for the use of a methodology that will allow us to build not only the lexical database but also the interfaces for its manipulation.

5.1 Preliminary Considerations

As we are interested in the creation of conceptual models for the creation of taxonomy based electronic dictionaries and lexical databases, before we describe the proposed conceptual model, it is necessary to make some precisions.

All the previous models only allow creating tree shaped taxonomies. The new model allows to represent DAG-shaped taxonomies where (as in the previous models) meaning is represented as a language independent entity; in particular, each meaning may constitute a category, and can be at the same time the extension of one or more categories (see figure 1). The meaning of words is thus represented using a taxonomic classification of concepts that are used as semantic primitives and where linguistic and conceptual knowledge are clearly separated (see figure 2).
5.2 Description of the Conceptual Model

In order to design the conceptual model, we used the E-R model, which is broadly used for database analysis, and as a unified framework for the developing of database systems. It allows us to identify the data objects and their relations using a graphical notation: it defines all the information that is introduced, produced and transformed in an application. It represents a network of data for a given system and it is useful in applications where the data and its relations are complex.

E-R models are based on several basic semantic elements: entity sets, relationship sets, attributes and values. Following the recommendations in (Pressman, 1997; Silberschatz, Korth and Sudarshan, 2001), entity sets are represented by rectangles, attributes by ellipses, relationship sets by rhombi, and roles label relationship lines.

An entity set represents an object of interest in the domain. The attributes of an entity set describe its properties using appropriate values. Relationship sets describe the interactions that may exist among one or more entity sets and connect the entity sets by means of lines.

Non-directed lines indicate that the cardinality is many to many (M:N). Lines are also used to link the attributes to the entity sets or relationship sets.
5.2.1 Entity Sets

In our conceptual model, the Meaning entity set from which the rest of entities depend upon is shown (see figure 3). The Term entity set represents all the terms in the terminological database. The Category entity set denotes the category to which each meaning belongs to. The entity Comment represents all the possible comments than can be associated with the terms.

5.2.2 Relationship Sets

The Synset relationship set between Meanings and Terms denotes the set of synonyms that belong to a meaning and it has a M:N cardinality because a set of synonyms can have several terms (synonymy) and the same term can be in different synonym sets (polysemy).

The See relationship set denotes the set of terms that, being not synonyms, are related to a synset and it has a M:N cardinality because one term can refer to others and it can also appear referenced by other terms.

The BelongsTo relationship set denotes the category which a meaning belongs to, and it has a M:N cardinality because there are several meanings that correspond to a category, and one meaning can be in several categories. This implies that our classification is not lexical (there is no direct relation between categories and terms) but semantic (terms are related to meanings, and meanings are related to categories).

The TermComment relationship set denotes the comments associated to each term and it has a M:N cardinality because one term can have several comments and the same comment can refer to several other terms.

The ParentOf relationship set denotes the conceptual taxonomy and it has a M:N cardinality because a category can have more than one parent category (except the root node) and a category can have several child categories. The Parent and Child roles label the lines, expressing the relationship between categories, and stating that a category is parent of a child category.

Note that the M:N cardinality of the ParentOf relationship is the big difference between the current proposal and the previous ones (Vaquero, Sáenz and Lopez, 2003).

5.2.3 Attributes

The Category entity set has the CategoryName attribute that denotes the textual name of the category. The Meaning entity set has the Definition attribute that denotes the textual definition of the meaning. The Term entity set has the TermName attribute that denotes the textual name of the term.
Finally, the Comment entity set has the CommentText attribute that denotes the text of the comment.

![Conceptual Model for a Monolingual Dictionaries](Image)

**Figure 3.** Conceptual Model for a Monolingual Dictionaries

6. LEARNING TOOLS

The basic functionality of the query and administration tools of monolingual dictionaries for language learning is present in the interface shown in figure 4 (following our previous work (Vaquero, Sáenz and Lopez, 2003). In particular, we highlight that our tools allow the visualization of possibly complex DAG-shaped taxonomies and how the constraints imposed on the lexical database help to build a dictionary following a constructivist approach. For the interested reader, see (Sáenz and Vaquero, 2005b) for the database design issues.

As it can be seen in figures 5 and 6, the new interfaces allow the visualization of complex taxonomies where a category can be the child of more than one category. The graph that is shown in the figures is a simple one, but it shows the functionality of this part of the tools. If a category has more than one parent category, as platypus, the tool warns the user and asks him to choose the path to follow: in this case oviparous or mammal. This could seem simple, but in more complex graphs it helps the user to see the part of the graph he is interested in.
The relational database design methodology is not only focused on representing data and their relations, but more important for us in this work, constraints about them. These constraints allow us to impose restrictions for both data and relations that any database instance must obey. However, because of the authoring nature of lexical databases, we cannot impose all of the identified constraints (since there is absent information which can be known afterwards). Therefore, we are ought to provide consistency checking features to the lexical database authors. These features must inform the author about authoring constraints which are violated by the instance database currently being constructed by him. Such constraints which may be violated during the authoring process are known as soft constraints, by contrast with the hard constraints that every database instance must hold at any time.

Our tools do include checking soft constraints in the 'Inconsistency Checking' ('Informe de Inconsistencias') tab. 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 among terms (polysemy, synonymy) (Vaquero, Sáenz and Barco, 2001). For instance, the new tool gives hints for detecting circular definitions and possible lacks of synonyms and related terms.

![Figure 4. Interface for Monolingual Dictionaries](image-url)
Figure 5. Visualization of the DAG-shaped Taxonomy with Multiple Inheritance

Figure 6. Subgraph for the Platypus Category
7. CONCLUSIONS AND FUTURE WORK

From the proposed conceptual model, we have created and tested tools that augment the capabilities of the previously developed ones. With these tools, the semantic relations implicated in simple and multiple inheritance situations can be learned, because the dictionaries that can be constructed are based on a DAG-shaped taxonomy with a single relationship. The dictionaries created with this tools explicitly represent the conceptual knowledge that is normally absent in most electronic dictionaries. With these tools, it is easy to promote team work and collaborative learning.

As for future work, first the tools must be installed in the classroom and then test them by creating and querying dictionaries just to be sure that the foreseen pedagogical objectives are fulfilled. Future versions of the tools must include more linguistic concepts that can be useful for the comprehension and mastery of languages. A long-term goal as a consequence of the evolution of this work is the integration of linguistic resources that nowadays are built heterogeneously in several places. Our more ambitious goal is to study, for a given domain, the overlapping phenomenon over multiple relations, beginning with the “IS A” and “PART OF” ones. Finally, a collaborative methodology may be defined and implemented in our tools.

REFERENCES


Norman, K., 1994, Navigating the educational space with HyperCourseware, Hypermedia, Vol. 6, pp. 35-60.


Raguenaud, C., 2000, Managing complex taxonomic data in an object oriented database, In proceedings of the EDBT 2000 PhD workshop, March 31-April 1, Konstanz, Germany.


Smith, B., Ceusters, W., Temmerman, R., 2005, Wüsteria, Accepted for MIE 2005, Geneva, 28-31 August.


