BaDELE3000: An implementation of the lexical inheritance principle

(1) María A. Barrios, (2) María Socorro Bernardos
(1) Universidad Complutense de Madrid
auxiba@filol.ucm.es
(2) Universidad Politécnica de Madrid
sbernardos@fi.upm.es

Abstract

This paper presents a database, called BaDELE3000, developed for a computer application for learning Spanish called CALLEX-ESP. Emphasis is placed on the potential properties of Lexical Functions in order to get a set of nouns that share identical values for collocations. The database implements the lexical inheritance principle which together with glosses of lexical functions makes it possible, for different sets of words with identical semantic labels, to inherit some collocations automatically. Thanks to this process, productivity and coherence of the final result is guaranteed.

Keywords

Lexical Inheritance Principle, Lexical Function, collocation, Meaning-Text Theory, lexical database.

1 Introduction

In 2002 the Institute for Information Transmission Problems of the Russian Academy of Sciences created CALLEX (Apresjan et al., 2002, 2003), a computer-assisted language learning tool for Russian, English and German. CALLEX allows the learners to extend their lexical knowledge by means of five types of games. Although this program is not directly linked to the Explicative and Combinatory Lexicology, the use of LFs allows it to be considered an implementation of the MTT. The systematisation of the system shows semantic and syntactic similarities and differences among LFs of the same family and their relation with lexemes that belong to the same lexical field (see collocations listed by different LFs in Apresjan et al. 2007). The adaptation of that tool for Spanish, called CALLEX-ESP, is under development in the phase of encoding the lexical relations; when concluded, the software system will produce many exercises automatically, although by the final version of CALLEX-ESP the model of DiCE (see footnote 4) exercises for Spanish could be taken in account.

This paper deals with a first version of a resource created for CALLEX-ESP: a database called BaDELE3000 (Base de Datos para el Español como Lengua Extranjera). This database contains information about the 3,000 most frequently used nouns in Spanish, a lexicon that was considered useful enough for an intermediate-advanced (learning) level here. BaDELE3000 has been designed having in mind three main aims:
1. To make each piece of relevant information easily accessible, either alone or in relation to others

2. To make the addition of information in the database as automatic as possible

3. To implement a principle of the MTT, the Lexical Inheritance Principle (Mel’čuk and Wanner, 1996; Mel’čuk, 1996: 76-78).

In order to accomplish the first aim, a systematic development process was followed, based on general principles and guidelines for the construction of data models and relational databases. To achieve the second and the third aims, Lexical Functions were used and the Lexical Inheritance Principle was applied. This principle states that the keywords (nouns) of some noun-verb collocations share some semantic properties. The use of this principle lets us automatically get several Spanish syntagmatic relations – i.e. values of Lexical Functions (LFs) – for some lexical frames. These lists of word combinations, grouped by LFs, are required by CALLEX.

The purpose of this paper is to show how this was done. Section 2 is a brief introduction to LFs and the Lexical Inheritance Principle; Section 3 describes the main components of the database; and Section 4 focuses on the levels of inheritance that the database supports and offers some results. Finally conclusions are drawn.

2 The Principle of Lexical Inheritance

Pustejovsky’s generative lexicon (1995) and WordNet’s lexical database\(^1\) are based on semantic inheritance, i.e. the semantic classification of lexemes (for example, *cat*, *dog* and *horse*) into a class (in this case, ANIMAL) which includes all the semantic features they share. Mel’čuk and Wanner (1996) apply a different approach to the lexicographic representation and propose the Lexical Inheritance Principle\(^2\), which they implemented and validated for the lexical field of emotion in German. As these authors suggest, in order to prove the existence of such principle, the semantic features of words that share some values for certain syntagmatic LFs can be analysed. This analysis can help to find generalizations regarding their restricted lexical co-occurrence. Thus, it will be possible to foretell that the verb *éprouver* (to feel) combines with every lexical unit meaning ‘emotion’: “Lexemes with common restricted co-occurrence also share semantic features (…). It must be possible to generalize restricted lexical co-occurrence instantiations along semantic lines” (Mel’čuk and Wanner, 1996: 209).

The authors studied the semantic features of the words that mean GEFÜHL (emotion), such as *Freude* (joy), *Scham* (shame), *Begeisterung* (enthusiasm), etc., and confirmed that some collocations are common for all the nouns of this lexical field, for instance *show* (joy, enthusiasm, hatred, etc.). Rather than repeating those collocations for each lexicographic entry, Mel’čuk and Wanner proposed two kinds of entries: a private one and a public one. The

\(^1\) [http://wordnet.princeton.edu/](http://wordnet.princeton.edu/)

\(^2\) Although the principle of lexical inheritance had already been applied to some entries of the DECFC (cf., DECFC vol. IV, p. 166, the entry *carotte* relating to legume), Mel’čuk and Wanner (1996) were the ones who introduced it.
private entry captures the collocations that are specific for the noun involved in that entry and the public entry captures the collocations that are common among all the nouns belonging to the lexical field involved by that entry.

Sanromán (2003:89) proved that collocations are at least partially semantically motivated; she employed the lexical inheritance principle applied in the DiCE (Alonso Ramos and Sanromán: 2000), which is a database based on the MTT principles containing the lexicon of emotion – fourteen lexical entries so far. Both authors followed Mel’čuk and Wanner’s methodology and applied it to the same lexical field, viz. emotions, in Spanish. The DiCE contains some shared values for some LFs, for instance, sentir (to feel) is a shared value for Oper1 for Spanish words meaning ‘feelings’.

Outside the framework of the MTT, such a principle has been implemented by Bosque (2006) in a certain way in the Diccionario Combinatorio Práctico del Español (Practical Combinatory Dictionary of Spanish) which contains common collocatives for many different entries grouped in 18 generic ones (e.g., 12 entries for the generic entry MONTH, 52 entries for LANGUAGE, 35 entries for COLOUR, etc.). For instance, the generic entry MONTH contains the participle entrado (in the sense of ‘almost in the middle of’) which is not present in month as a normal entry; that means that collocations as enero bien entrado (‘almost in the middle of January’) are possible but not *mes entrado (this sense should be ‘almost in the middle of the month’). On the other hand, the normal entry month contains the verb cumplir, which is not in the generic entry, because it is possible to say cumplir dos meses (to be two months old) but not *cumplir dos eneros (*to be two Januaries old).

Coming back to the MTT framework, the lexical inheritance principle can also be applied in a different way. Instead of studying the semantic features of the words belonging to a lexical field, it is possible to work with the lexemes that share identical values for the same LF: “The inheritance of lexical co-occurrence properties could be implemented using the same methodology: by grouping all the lexemes that have identical values of the same LFs into one class” (Mel’čuk and Wanner: 2006, 210-211). Furthermore, Mel’čuk (1996: 76-78) adds that the Lexical Inheritance Principle can be applied in two ways: the shared values of LFs can be generalized on the side of the keyword or on the side of the LF itself.

In this work, the latter alternative has been adopted: the lexical inheritance principle has been applied to a lexicon of 3,000 Spanish nouns, grouped according to semantic labels. A semantic label is the formalization of the meaning of the keyword (Polguère, 2003:43). Words under the same semantic label represent a lexical field. For instance, the semantic label accidente geográfico (geographical feature) includes lexical units such as río (river), mar (sea) and montaña (mountain). One of the designing guidelines of BaDELE3000 is the hypothesis that, at least, a LF could be applied to all the words with the same semantic label and that some value(s) of that LF could be also shared. This is the reason why the application of the lexical inheritance principle is very useful for our purpose.

For example, let us take a look at the LF called CausFunc0, which means ‘to cause something to begin to exist’. First, the semantic labels of the nouns that can be combined with a verb that

---

3 Chapter 3 contains a description of Spanish emotion nouns based on the concept of lexical inheritance.

means ‘to cause something to begin to exist’ are selected – for instance, *energía* (energy), *fenómeno meteorológico* (meteorological phenomenon), *sentimiento* (feeling) etc.; and then, they are related to the lexical units that express that sense – for instance, *producir* (to produce energy), *provocar* (to cause rain), *despertar* (to arouse a feeling) etc.

3 BaDELE3000

As mentioned before, the database we have developed has been called BaDELE3000 and contains the 3,000 most frequently used nouns in Iberian Spanish. This set of words is based on the statistical study of the Corpus Cumbre by Almela *et al.* (2005) and refined with the information provided by some native speakers.

First, a knowledge acquisition process (Scott *et al.*: 1991) was carried on in order to get a comprehensive understanding of the characteristics and requirements concerned in our matter. The techniques applied involved mainly the analysis of many publications (e.g., Wanner: 1996; Grossmann and Tutin: 2003), dictionaries (*Dicouèbe, DiCE, DECFC*) and the application *CALLEX*; and the exchange of information between the linguists and the computer engineers.

The following paragraphs describe the most important concepts and specifications educed from that knowledge acquisition process which are relevant for this paper.

3.1 Semantic labels

As explained above, the semantic label of a word is the formalization of the sense of the word, which is usually the immediate generic sense of the word. Each Lexical Unit has one and only one semantic label, but a semantic label can include several lexical units (which, consequently, belong to the same lexical field).

Semantic labels are hierarchically organized. They can be hyperonyms and/or hyponyms of other semantic labels (see Table 1).

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
<th>Lexical units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evento (event)</td>
<td>Evento fásico (phasic event)</td>
<td>Fenómeno (phenomenon)</td>
<td>Fenómeno natural (natural phenomenon)</td>
<td>alba, eco (dawn, echo)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Desastre natural (natural disaster)</td>
<td>inundación, terremoto, (flood, earthquake)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fenómeno meteorológico (meteorological phenomenon)</td>
<td>tormenta, viento, lluvia (storm, wind, rain)</td>
</tr>
</tbody>
</table>

Table 1: An example of the hierarchy of semantic labels and lexical units

One of the hypotheses of this work, as mentioned in section 2, is that lexical units classified under the same semantic label can share one or more LFs, and sometimes even some values of those LFs.

---

5 This is a corpus of 20 million words of written and spoken Spanish.
3.2 Lexical Units

Each sense of a word contained in the database corresponds to a lexical unit. Each lexical unit is classified into a single semantic label\(^6\) (see semantic labels above) and can be applied to one or several LFs (see LFs below), whose values are also lexical units\(^7\).

3.3 Lexical functions

A LF relates a lexical unit to one or several values, which are lexical units that express a specific meaning for that LF. These values are grouped according to their proximity.

LFs can consist of other LFs related by different binary operators (such as concatenation). The LFs defined by Mel’čuk (1996) are the starting set and non standard LFs can be added when necessary. LFs have been classified according to the LFs classification of Mel’čuk (1996): each LF belongs to a class. In addition, these classes are also structured and each one can be classified into other classes. LFs can be expressed in natural language by semantic definitions – one at most for a given LF – and/or glosses (see below). Since lexical units with the same semantic label share certain values of certain LFs, the LFs attached to this semantic label can be related to those values. That way the lexical units belonging to the same semantic label can inherit those values for those LFs. Sometimes no values are shared; and it can be only stated that certain LFs can be applied to certain semantic fields.

3.4 Semantic definitions and glosses

The meaning of a LF can be translated into one of several natural language expressions, which are called paraphrases or glosses (Kahane and Polguère, 2001; Alonso Ramos, 2005, 2006). Expressions that consist of a single lexical unit are especially relevant for our purposes because they are values of the LF they represent. They can be used as ‘default values’ of LFs (Alonso Ramos; 2005) and hence, they can be applied to lexical units under the same semantic label. That is why it is important to distinguish between glosses, which are single lexical units, and semantic definitions, which can take the form of a phrase. For instance, the semantic definition of \textit{CausFunc}_0 would be ‘to cause something to exist’, and glosses of that LF would be \textit{to make}, \textit{to create}, etc.

An entity/relationship diagram (Chen: 1976) that matches the specification above was made (see figure 1). It includes the main components (entities, relationships and attributes) of the database. The entities (rectangles) are the concepts with existence on their own – in this case, the headings of the subsections; the relationships (rhombus) reflect the connections among the entities; and the attributes (circles) represent their features or properties. This diagram was transformed into tables of a relational database (Codd, 1970). Lack of space prevents more details here; suffice it to say that the result accomplishes the normalization forms of relational databases.

---

\(6\) Lexical units can be organized according to many criteria. Future versions of the database will include other classifications. We intend to create an ontology that can be useful not only in the Computational Linguistic field, but also in other fields.

\(7\) Actually, the lexical units that are studied in depth are distinguished from the lexical units that are included only as values of LFs. Since most of the data for the first ones are not pertinent for this paper, this distinction has been dealt with here.
databases. Several views of the tables and their content have been created as well as some queries in order to see the particular information required at different moments. Section 4 presents some examples.

4 Methodology and Results

The lexical inheritance principle has been applied in two phases: in the first one, the principle is used at the semantic label level – i.e., values are inherited by hyponyms of the given semantic labels (see Table 2 and 3) – in the second one, it is employed at the lexical unit level (see table 4) – i.e., values are inherited by the lexical units belonging to the given semantic labels. In order to complete the information in the database, an additional phase is required where all data are included manually (see table 5 and 6). The following paragraphs describe the procedure followed in more detail.

First, in a non automatic phase: for each LF, a broad list of nouns that can be combined with an expression meaning the same as the LF – i.e. its semantic definition and/or gloss(es) – was made. For instance, Real1 means ‘to use something according to its destination’, so nouns such as disciplina (discipline), juguete (toy), herramienta (tool), deporte (sport), etc. are susceptible of combination with different glosses of that LF – e.g., estudiar [una disciplina] (to study a discipline), jugar [con un juguete] (to play with a toy), usar [una herramienta] (to use a tool), limpiar [con un producto de limpieza] (to clean with a cleaning product), circular [por una via] (to move about a road), cumplir [una ley] (to obey a law), practicar [un deporte] (to practice a sport).
The semantic labels of the nouns matching that condition were selected, and they were related to the verbal glosses expressing that sense. Table 2 shows some examples of this for Real₁. The first column is the LF; the second is the semantic label, and the third is a gloss of the LF that can be applied to the semantic label.

<table>
<thead>
<tr>
<th>LF</th>
<th>Semantic label</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real₁</td>
<td>Ropa y complementos (clothes and accessories)</td>
<td>llevar (to wear)</td>
</tr>
<tr>
<td>Real₁</td>
<td>Disciplina (discipline)</td>
<td>estudiar (to study)</td>
</tr>
<tr>
<td>Real₁</td>
<td>Juguete (toy)</td>
<td>jugar (to play)</td>
</tr>
<tr>
<td>Real₁</td>
<td>Herramienta (tool)</td>
<td>usar (to use)</td>
</tr>
<tr>
<td>Real₁</td>
<td>Deporte (sport)</td>
<td>practicar (to practice)</td>
</tr>
</tbody>
</table>

Table 2: Excerpt from the ascription of the glosses of the LF Real₁ to the semantic labels

Then the semantic labels classified under those first selected ones would inherit their glosses automatically. Table 3 displays the result of this inheritance phase for Real₁ and the semantic labels classified as ‘clothes and accessories’. The forth column shows if the value has been automatically ascribed – i.e. inherited – and the fifth one if the inherited value has been rejected after having been examined by an expert. The database contains 219 combinations of glosses and semantic labels for Real₁, of which nearly half (105) have been inherited.

<table>
<thead>
<tr>
<th>LF</th>
<th>Semantic label</th>
<th>Gloss</th>
<th>Inherited</th>
<th>Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real₁</td>
<td>Ropa (clothes)</td>
<td>llevar (to wear)</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Real₁</td>
<td>Complemento (accessory)</td>
<td>llevar (to wear)</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Real₁</td>
<td>Calzado (footwear)</td>
<td>llevar (to wear)</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

Table 3: Excerpt from the inheritance of the gloss within the semantic label hierarchy

The second phase of the application of the lexical inheritance principle lets the glosses reach the lexical units under the semantic labels previously handled. For Real₁, 2,373 collocations have been obtained; some of them are shown in Table 4. The more prolific a LF is and the more shared values for lexical units it produces, the more useful is the application of the lexical inheritance principle.

<table>
<thead>
<tr>
<th>LF</th>
<th>Semantic label</th>
<th>Gloss</th>
<th>Lexeme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real₁</td>
<td>Ropa (clothes)</td>
<td>llevar (to wear)</td>
<td>pantalones (trousers)</td>
</tr>
<tr>
<td>Real₁</td>
<td>Complemento (accessory)</td>
<td>llevar (to wear)</td>
<td>guantes (gloves)</td>
</tr>
<tr>
<td>Real₁</td>
<td>Calzado (footwear)</td>
<td>llevar (to wear)</td>
<td>sandalias (sandals)</td>
</tr>
<tr>
<td>Real₁</td>
<td>Instrumento musical (musical instrument)</td>
<td>tocar (to play)</td>
<td>flauta (flute)</td>
</tr>
<tr>
<td>Real₁</td>
<td>Herramienta (tool)</td>
<td>usar (to use)</td>
<td>tijeras (scissors)</td>
</tr>
<tr>
<td>Real₁</td>
<td>Utensilio de limpieza (cleaning utensil)</td>
<td>usar (to use)</td>
<td>escoba (broom)</td>
</tr>
<tr>
<td>Real₁</td>
<td>Producto de higiene (care product)</td>
<td>usar (to use)</td>
<td>gel (gel)</td>
</tr>
</tbody>
</table>

Table 4: Excerpt from the inheritance of the glosses from the semantic labels to their lexical units

Glosses are a good way to obtain possible collocations, but there may be some values that need to be added manually. For instance, cortar con las tijeras (to cut with some scissors) o pasar la escoba (to brush) for files 5 and 6 of Table 4. Moreover, when collocations are not
due to any semantic motivation, they cannot be automatically included in the database. That is
the reason why two Spanish combinatory dictionaries (Bosque: 2004; 2006) were very useful
in completing the task. Typical cases belong to one of the two following categories: (i) there
are lexemes which combine with a gloss for no semantic reason (see table 5 for examples with
glosses of $\text{IncepFunc}_0$, ‘to start to exist’, for lexical units found in Bosque’s entry (2006:
886) $\text{nacer}$), where support verb lexical entries (Reuther: 1996) are remarkably useful; (ii)
there are more values for the LF of a lexical unit and these new values can be considered
synonyms of the (inherited) glosses (see table 6 for examples of inherited and added values
for the LF $\text{CausFunc}_0$ of lexical units whose semantic label is ‘fenómeno meteorológico’
(meteorological phenomenon): 22 combinations, 8 inherited from their semantic label and 16
added manually).

<table>
<thead>
<tr>
<th>LF</th>
<th>Semantic label</th>
<th>Lexeme</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{IncepFunc}_0$</td>
<td>Parte del cuerpo humano (part of the human body)</td>
<td>pelo (hair)</td>
<td>$\text{nacer}$ (to grow)</td>
</tr>
<tr>
<td>$\text{IncepFunc}_0$</td>
<td>Parte del cuerpo animal (part of the animal body)</td>
<td>pluma (feather)</td>
<td>$\text{nacer}$ (to grow)</td>
</tr>
</tbody>
</table>

Table 5: Excerpt from glosses added

<table>
<thead>
<tr>
<th>CausFunc$_0$</th>
<th>Lexeme</th>
<th>Gloss</th>
<th>Added values</th>
</tr>
</thead>
<tbody>
<tr>
<td>lluvia (rain)</td>
<td>provocar (to cause)</td>
<td>originar, causar (to originate), producir (to produce)</td>
<td></td>
</tr>
<tr>
<td>chubasco (heavy shower)</td>
<td>provocar (to cause)</td>
<td>originar, causar (to originate)</td>
<td></td>
</tr>
<tr>
<td>tormenta (storm)</td>
<td>provocar (to cause)</td>
<td>ocasionar, desencadenar, desatar (to break)</td>
<td></td>
</tr>
<tr>
<td>precipitación (precipitation)</td>
<td>provocar (to cause)</td>
<td>dar lugar a (~-s) (to give rise to)</td>
<td></td>
</tr>
<tr>
<td>viento (wind)</td>
<td>provocar (to cause)</td>
<td>levantar (to start blowing)</td>
<td></td>
</tr>
<tr>
<td>nieve (snow)</td>
<td>provocar (to cause)</td>
<td>causar (to produce)</td>
<td></td>
</tr>
<tr>
<td>temporal (big storm)</td>
<td>provocar (to cause)</td>
<td>ocasionar, desencadenar, desatar (to break)</td>
<td></td>
</tr>
<tr>
<td>nube (cloud)</td>
<td>provocar (to cause)</td>
<td>originar (to originate), formar (to develop)</td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Excerpt from the glosses and values of the LFs of the lexical units

So far around 8,676 Spanish collocations have been automatically obtained. After adding the
synonyms for the glosses and the collocations not motivated by the semantics, the database
has increased substantially and contains 20,406 collocations currently.

5 Conclusions

This paper explains how the lexical inheritance principle can be useful for developing a
lexical resource (in particular, a database). Even though this is an ongoing work, its results so
far show that the application of this principle supports the automatic addition of some
collocations that are repeated for lexical units belonging to the same lexical field, thus
reducing the time and the effort the developers and the lexicographers have to devote to this
task. This is especially relevant for applications that do not search for an exhaustive
description of a language, but for the formalization of the lexical relations that are most
frequently used.
Although BaDELE3000 was created for the CALLEX-ESP project, it is a self-contained resource that can be used for different applications and purposes. For instance, a database such as this can help disambiguate a term by examining the words that are close to it and comparing them to the collocations of that term in the database. It can also be employed to find the right word/verb when generating a text in Spanish.

Acknowledgements

We would like to show our gratitude to Ignacio Bosque for his comments on the second version of this paper; to Igor Mel'čuk, Alain Polguère and Sylvain Kahane for their generosity in answering all our questions; and especially to Igor Boguslavsky for his support on CALLEX-ESP tasks and to Juri Apresjan because his systematisation of CALLEX lead us to develop this work. We apologize for the mistakes that the reader can find: we are the only one responsible for them.

Bibliography


