EDBL: a Multi-Purposed Lexical Support for the Treatment of Basque

Aduriz I., Aldezabal I., Ansa O., Artola X., Díaz de Ilarraza A., Insausti J. M.
IXA Taldea
Department of Computer Languages and Systems
University of the Basque Country
20080 Donostia - The Basque Country (Spain)
[jibxuxen@si.ehu.es]

Abstract

EDBL (Euskararen Datu-Base Lexikala) is a lexical database (LDB) for Basque. Seen as a large repository of lexical information, it acts as the basis for different tasks in automatic processing: it must be both source and support for the lexicons needed in different applications. Besides providing for multiple applications, the lexical database is intended to be neutral in relation to the different linguistic formalisms, flexible and open enough to accept new types of information, and easy to use.

The applications of the database are presently the following: morphological analysis, spell checking and correction, (semi)automatic lemmatisation and tagging; syntactic analysis and analysis of textual corpora will be tasks we will inplement in the short term. These applications constitute the necessary foundations of present and further work on the processing of Basque.

This paper presents the conceptual schema and the main features of this database, conceived as a general lexical basis for the automatic treatment of Basque. Extended Entity Relationship diagrams and TEI-conformant feature-structures coded in SGML are used to explain the conceptual schema of the LDB. In the short term, these feature-structures will be used as a standard input and delivery format for the tools described above.

Introduction

In this article we introduce the Lexical Database for Basque (EDBL), which is currently being used as a lexical support for the automatic treatment of Basque. EDBL (Agirre *et al.*, 1995), a large source of lexical information (with about 70.000 entries), has been designed as a *multi-purposed* lexical support for different processes: a goal-independent lexical source for the treatment of the Basque language. The lexicons obtained from the database are afterwards used in tools such as the morphological analyser (Urkia, 1997), the speller/checker (Agirre *et al.*, 1992) as well as the tagger/lemmatiser (Aduriz *et al.*, 1996). In the short term, we will use TEI-conformant feature-structures coded in SGML as a standard input and delivery format for the tools described above.

Apart from the features already described, EDBL is characterised by the following ones:

- *Neutral*: the linguistic descriptions held in it should not constrain any applications in the future. This does not mean, obviously, that no formalism will be used in these linguistic descriptions, but that the LDB will remain open to new descriptions, compatible or not with previous ones. We have used the well-known "Two-level morphology" (Koskenniemi, 1983) for the morphological description of Basque.
- *Open and flexible*: EDBL is capable of being easily adapted when necessary, as well as of accepting new kind of data.
- *User-friendly*: This database was originally designed and adapted to be used by applications and human (specialised or not) users. Applying *user-friendliness* to its interface, the database was conceived to help both normal and specialised users, giving them strong support when the tasks can be automatically managed.

First of all, we shall introduce the structure of the database, describing also the reasons for which we have adopted this structure. Later on, we will describe the TEI-conformant SGML Feature Structures representation we have developed for data in EDBL.

General Description of the Database Structure

In order to describe the structure of this database, we have used the Extended Entity Relationship (EER) data model based on the Entity Relationship (ER) model, as we consider it suitable for describing the hierarchical relationships amongst the different objects in EDBL.

Main Entities in EDBL

The main data entity in EDBL is the one called EDBL Units, the key of which is composed by a headword and a homograph identifier, as in any conventional dictionary. This homograph identifier lets us distinguish different senses for a dictionary entry. This data class can be viewed from three different standpoints, as it holds three different specialisation relationships depending on several features we describe below. Although all the specialisations are total (all entries belong to the three specialisation units, thick line in the diagram), we want to remark that each one is separately followed by a disjointed specialisation (d within a circle). Let us have a glance at them.

Standard and Non-Standard Forms.

This partial specialisation divides the considered elements into standard and non-standard entries (figure 1). For any lexical entry in the database, being a standard element implies it is correctly spelled and hence, it is accepted by The Basque Language Academy as a lexical entry for the Standard Basque. As the normalisation of the language is still in process, the number of non-standard forms currently used is quite large. The forms belonging to the

Non-Standard class can be either dialectal forms (both at a lexical and a morphemic level) or badly used words. The relationship placed between Standard and Non-Standard elements allows us to relate those correct forms to the ones considered to be incorrect. The existence of more than one dialectal form for any standard entry implies that the cardinality of this relationship is 0:n, i.e. we can relate zero or more Non-Standard entries to a Standard one in our database.

Dictionary Entries and Other Entries

The main specialisation in this database is the one that separates Dictionary Entries from the Other Entries (figure 2). In the class related to dictionary entries, we include any lexical entry that could be found in an ordinary dictionary, those with a main category such as nouns, verbs, adjectives, etc. as well as abbreviations and initials, compounds and derivatives (figure 3). On the other hand, the class named Other Entries is specialised in two subclasses, Non-Independent Morphemes and Inflected Forms. While non-independent morphemes require a lemma for their use inside a word, inflected forms implicitly take case, number, etc. and hence need no morphemes to add such information. We have created a relationship between all inflected forms and their corresponding lemma entries. The cardinality of the relationship is as follows: while every inflected form has only one dictionary entry lemma (cardinality 1:1), every lemma (a dictionary entry) may be related to zero, one or more inflected forms. As we can see in figure 2, both subclasses (Non-Independent Morphemes and Inflected Forms) are the total disjunctive specialisation of the Other Entries class.

Figure 1: Standard and Non-Standard Entities´ Relationship

Figure 2: Dictionary Entries and Other Entries

The entries we have considered as inflected forms are either the ones that would need a very complex morphotactic treatment (e.g. inflected verb forms), or the ones that cannot be morphologically analysed and decomposed

(other inflected forms), as the whole morphological information is conveyed by themselves.

Multiword and Simple-Word Entries.

In order to finish this brief description, we will explain the specialisation of the main class as divided into two partial subclasses, one containing Simple Lexical Units (SLU) and another one containing Multiword Lexical Units (MWLU); look at figure 4. We define a SLU as any string of characters between two blanks, that is, a single typographic unit (hyphened forms included). On the other hand, we consider MWLU all lexical units expressed by two or more items placed between blanks. The way they are stored depends upon whether they are SLU or MWLU, as the classification of the MWLU entries is quite complex. Let us have a look at each one of them in turn.

Figure 3: Dictionary Entries

Figure 4: SLU and MWLU Entries

Simple Word Entries' morphological components (figure 5). As Basque is an agglutinative language, it presents a relatively high power to generate inflected word-forms; any word entry independently takes each of the necessary elements (the affixes corresponding to the determiner, number and declension features) for the different (syntactic case included) functions. Moreover, noun ellipsis can occur inside a complex noun due to recursive constructions, even though, in practice, it is not usual to find more than two levels of recursion in a word-form.

Therefore, it is necessary to apply a mechanism that lets us know the morphotactic formation of the morphemes in a simple word –that is, the way the morphemes are linked–, and their necessary morphosyntactic information. "The Two-level Morphology" proposed by Koskenniemi (Koskenniemi, 1983) was chosen as the most adequate to represent lexical information in Basque, due to the reasons we mention next: it distinguishes the algorithm and the language knowledge; it is a general model (applicable to any language); it works either as a morphological analyser or generator; and finally, it distinguishes both the surface and the lexical level.

We use the Morphotactics relationship in order to explain the different morphological aspects SLU-s adopt. The different entities that take part in this relationship are the following: Single-word entries, Two-level forms of the entries, Continuation Classes and Lexicons. Every entry is at least related to an only Two-level form, but it may be related to even more entries (Two-level forms usually have diacritics by which we the proper morpho-phonological rules will be selected). Every Two-level entry is related to a Lexicon, as well as being attached to a Continuation Class. By relating Two-level forms to Continuation Classes, we define the set of morphemes we can add or agglutinate after the stem of an entry.

Figure 5: SLU´s Morphotactical Relationship

Our system consists of a lexicon made up of about 70,000 items, grouped into 120 sublexicons and a set of 24 morpho-phonological rules that describe the changes occurring between the lexical and the surface level. We have only applied this method to SLU-s in Basque, as MWLU-s behave differently: we explain this behaviour upon the use of some MWLU realisation patterns described below.

Multiword Entries' morphological components (figure 6) It's not always easy to decide whether an item must be defined as a MWLU; nonetheless, it is obvious they show a very complex variety when constructed. First, MWLUs' components can appear contiguous or dispersed; in other words, they do not necessarily occur one after another; second, the order is not fixed, as some MWLUs must be formed in a restricted word order while others must not: an item may appear in different locations in the MWLU; third, the components may either be inflected or appear as an invariable form. In the case their components are inflected, some of them may accept any inflection whilst others must only take a restricted set of inflected attributes. Thus, we must define the restrictions for the components to accept a few inflected forms. And finally, some MWLUs are sure and some are ambiguous: we cannot accurately assure the same group of words to be a multiword entry in any context.

Figure 6: MWLU´s composition and morphological realisation

Following these features, we use a formal description where different MWLU realisation patterns are applied to each MWLU entry. As well as declaring the certainty of a MWLU, the realisation patterns restrict the possible order, contiguity and inflection rules (for each order) every element may adopt in the whole lexical unit.

EDBL entities represented as Feature Structures.

As it has been pointed out before, TEI-conformant SGML is used as delivery format for EDBL entities. In order to represent the structure of each one of the entities, typed Feature Structures (FS) will be used. Feature structures are very adequate to encode linguistic information, there is a well-developed theoretical framework for them, and it seems that their applicability to encode the information found in dictionaries, or in lexical databases for NLP, as is our case, is quite natural (Ide *et al.*, 1993).

Instead of defining our own formalism for FS's, we have adopted the one defined by TEI-P3, as we found it useful and neat for our purposes. Following TEI-P3, Feature Structure Declarations (FSD's) have been made for all the types of FS's that are used in EDBL. These FSD's reflect the hierarchic class structure of the database, and feature inheritance is used in order to make the definition of each class more consistent and comfortable (the attribute Base-Type is used in the definition of a type to declare the superclass or basic type from which the type defined inherits features).

Each one of the classes in the figures above define a different FS type. The main class defines the most general structure —`EDBL-Unit-FS`—, whose features are inherited by every class and every instance in the database. Let us show the FSD of this main class:

```
<fsdecl type="EDBL-Unit-FS">
 <fsdescr>Main class of EDBL: fetures that
       belong to every EDBL unit
 </fsdescr>
 <fdecl name="Key">
  <fdescr>Main key of every unit</fdescr>
  <vrange>&Key-Type</vrange>
 </fdecl>
 &Part-of-Speech
</fsdecl>
```

In the declaration above, `&Key-Type` and `&Part-of-Speech` are SGML entities that are defined elsewhere. The features in the FSD's may contain different types of values specified by the element `vrange`. Let us show now the `Key` FS type, defined within the `&Key-Type entity:`

```
<fsdecl type="Key-Type">
 <fsdescr>Main key of every EDBL unit
 </fsdescr>
 <fdecl name="Headword">
  <fdescr>Entry word</fdescr>
   <vrange><str rel="ne"></str></vrange>
 </fdecl>
 <fdecl name="Homographe-Id">
  <fdescr>Homographe identifier used to
       distinguish two units with the
      same headword
  </fdescr>
  <vrange>
   <nbr value="0" rel="ge" type="int">
  </vrange>
 </fdecl>
</fsdecl>
```

Finally, let us show an example of an instance of EDBL, such as it will be delivered. It is the case of a derived noun, *berrerabilgarritasun*, Basque for *reusability*. The morphosyntactic information of a derived noun includes the set of features defined for the nouns —mainly subcategorization, and a series of tags indicating whether it is an animate, countable, measurable or mostly used in plural—, and the information depicting it as a derived word, that is, the cross-references to its base, and eventually to the prefixes and/or to the suffixes that are present in the word (one prefix and two suffixes in the example).

```
<p>
 <!-- berr+erabil+garri+tasun,
    lexicalised derived noun entry -->
 <fs type="Derivative-Nouns">
  <f name="Key">
   <fs type="Key-Type">
    <f name="Headword">
      <str>berrerabilgarritasun</str>
    </f>
    <f name="Homographe-Id">
     <nbr value="1"></f>
   </fs>
  </f>
  <f name="POS"><sym value="NOUN"></f>
   <f name="Morphology" org="set">
    <fs type="Morphology-FS">
    <f name="TWOL-Form">
      <str>berrerabilgarritasun</str>
    </f>
     <f name="Continuation-Class">
      <str>I</str></f>
     <f name="Sublexicon">
      <str>nouns</str></f>
   </fs>
  </f>
  <f name="Source"><sym value="IH"></f>
  <f name="Source-Form">
    <str>berrerabilgarritasun</str>
  </f>
  <f name="Noun-Features">
   <fs type="Noun-Features-FS">
    <f name="Subcategorisation">
     <sym value="Common"></f>
    <f name="Animate"><minus></f>
    <f name="Countable"><minus></f>
    <f name="Mesurable"><plus></f>
    <f name="Plural"><minus></f>
   </fs>
  </f>
  <f name="Derivation-Features">
   <fs type="Derivation-Features-FS">
    <f name="Base">
     <fs type="Key-Type">
      <f name="Headword">
       <str>erabili</str>
```

```
        </f> <!-- to use -->
        <f name="Homographe-Id">
          <nbr value="1"></f>
        </fs>
      </f>
      <f name="Prefix-List" org="list">
       <fs type="Key-Type">
        <f name="Headword">
          <str>ber</str>
        </f><!-- re- prefix -->
        <f name="Homographe-Id">
          <nbr value="1"></f>
        </fs>
      </f>
      <f name="Suffix-List" org="list">
       <fs type="Key-Type">
        <f name="Headword">
          <str>garri</str>
        </f><!-- -ble suf.-->
        <f name="Homographe-Id">
          <nbr value="1"></f>
        </fs>
       <fs type="Key-Type">
        <f name="Headword">
          <str>tasun</str>
        </f><!-- -ity suf. -->
        <f name="Homographe-Id">
          <nbr value="1"></f>
        </fs>
      </f>
    </fs>
  </f>
 </fs>
</p>
```

Conclusions

We have presented the EDBL database, a flexible and multi-purposed lexical support for the automatic treatment of the Basque language. The description of its structure is based on the EER data model and it will provide an SGML-structured input to all the tools using the database. As described in (Ide *et al.*, 1992, Le Maitre *et al.*, 1993), the data models applied in lexical databases are not adequate to fully capture the structural properties of lexical information The use of inheritance provided by object oriented (OO) databases would improve our data representation. Features like derivation and semantics would be better captured using the OO approach and that's why, in the short term, we plan to follow it for new versions of EDBL.

Acknowledgements

This research was supported by the Basque Government, the University of the Basque Country and the Interministerial Commission for Science and Technology (CICYT).

References

Aduriz I., Aldezabal I., Alegria I., Urizar R. (1996). EUSLEM: A lemmatiser/tagger for Basque. EURALEX'96, Gotteborg (Sweden).

Aduriz I., Aldezabal I., Artola X., Ezeiza N., Urizar R. (1996). Multiword Lexical Units in EUSLEM, a lemmatiser/tagger for Basque. Papers in *Computational Lexicographics*, COMPLEX´96, Linguistics Institute, Hungarian Academy of Science. Budapest.

Aduriz I., Alegria I., Artola X., Ezeiza N., Sarasola K., Urkia M. (1997). A Spelling Corrector for Basque Based on Morphology. *Literary and Linguistic Computing 12*, N. 1.

Agirre E., Arregi X., Arriola J. M., Artola X., Díaz de Ilarraza A., Insausti J. M., Sarasola K. (1995). Different issues in the design of a general-purposed Lexical Database for Basque. First International *Workshop on Applications of Natural Language to Databases*, NLDB'95. Versailles.

Agirre E., Arregi X., Arriola J. M., Artola X., Insausti J. M. (1994). Euskararen Datu-Base Lexikala (EDBL). Dept. of Computer Languages and Systems, Univ. of Basque Country (EHU/UPV). Techn. Report UPV/EHU/LSI/TR 8-94.

Alegria I., Artola X., Sarasola K., Urkia M. (1996). Automatic Morphological Analysis of Basque. *Literary and Linguistic Computing 11*, N. 4.

Ide N., Le Maitre J., Véronis J. (1993). Outline of a model for lexical databases. *Information Processing and Management*, 29 (2) (pp. 159--186).

Ide N. & Véronis J. (1995). *Text Encoding Initiative. Background and Context.* Kluwer Academic: Dordrecht.

Koskenniemi K. (1983). Two-level morphology: A General Computational Model for Word-Form Recognition and Production. University of Helsinki, Department of General Linguistics. Publications nº 11.

Le Maitre J., Ide N., Véronis J. (1993). Deux modèles pour la représentation des données lexicales et leur implementation orientée-objet. Centre Nationale de Recherche Scientifique, France.

Sperberg-McQueen C.M. & Burnand L. (1994) *Guidelines for Electronic Text Encoding and Interchange*. Chicago & Oxford.

Urkia M. (1997). Euskal morfologiaren tratamendu informatikorantz. PhD thesis, University of The Basque Country.