EDBL: a General Lexical Basis for the Automatic Processing of Basque

I. Aldezabal, O. Ansa, B. Arrieta, X. Artola, A. Ezeiza,
G. Hernández, M. Lersundi

Dept. of Computer Languages and Systems
Faculty of Computer Science (UPV/EHU), 649 p.k.
20080 Donostia (The Basque Country)
jiparzux@si.ehu.es

Abstract

EDBL (Euskararen Datu-Base Lexikala) is a general-purpose lexical database used in Basque text-processing tasks. It is a large repository of lexical knowledge (currently around 80,000 entries) that acts as basis and support in a number of different NLP tasks, thus providing lexical information for several language tools: morphological analysis, spell checking and correction, lemmatization and tagging, syntactic analysis, and so on. It has been designed to be neutral in relation to the different linguistic formalisms, and flexible and open enough to accept new types of information. A browser-based user interface makes the job of consulting the database, correcting and updating entries, adding new ones, etc. easy to the lexicographer.

The paper presents the conceptual schema and the main features of the database, along with some problems encountered in its design and implementation in a commercial DBMS[1]. Given the diversity of the lexical entities and the complex relationships existing among them, three total specializations have been defined under the main class of the hierarchy that represents the conceptual schema. The first one divides all the entries in EDBL into Basque standard and non-standard entries. The second divides the units in the database into dictionary entries (classified into the different parts-of-speech) and other entries (mainly non-independent morphemes and irregularly inflected forms). Finally, another total specialization has been established between single-word entries and multiword lexical units; this permits us to describe the morphotactics of single-word entries, and the constitution and surface realization schemas of multiword lexical units.

A hierarchy of typed feature structures (FS) has been designed to map the entities and relationships in the database conceptual schema. The FSs are coded in TEI-conformant SGML, and Feature Structure Declarations (FSD) have been made for all the types of the hierarchy. Feature structures are used as a delivery format to export the lexical information from the database. The information coded in this way is subsequently used as input by the different language analysis tools.

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 the language.

EDBL (Agirre , 1995; Aduriz , 1998) is a large store of lexical information that nowadays contains more than 80,000 entries, which has been conceived as a lexical basis, i.e. a goal-independent resource for the processing of the language. The lexicons obtained from the database are subsequently used in tools such as a morphological analyzer (Urkia, 1997), a spelling checker (Aduriz , 1997), a tagger/lemmatizer (Aduriz , 1996), etc.

EDBL is integrated in the chain of Basque processing resources and tools, and the information contained in it is exported when

[1] Oracle 8.

needed to be used as input by the language analysis tools.

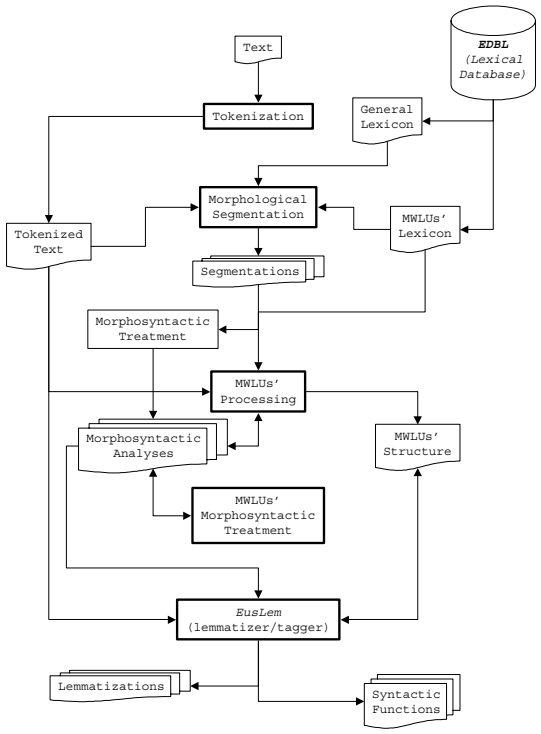

Figure 1: EDBL within the stream of language processing tools.

Apart from the features already mentioned, EDBL can be characterized as:

- 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 lexical database remains open to new descriptions. For instance, we are currently using the two-level morphology formalism, defined by Koskenniemi (1983), to describe the language morphology.
- Open and flexible.
- User-friendly: this database was originally conceived and designed to be used by applications and humans (specialized or not). A recently implemented browser-based interface makes the interaction with the database when querying and/or updating it easy to the lexicographers, and allows non-specialized users to consult the information contained in EDBL by means of a standard web browser[2].

In the following, we shall first introduce the database placing it in the chain of language processing tools developed in the group[3]. In Section 2, the conceptual schema of the database will be explained, using for that Extended Entity-Relationship diagrams and describing the rationale under this schema. Section 3 depicts very shortly the linguistic contents of EDBL, giving some figures. Section 4 is devoted to describing the TEI-conformant feature structures-based representation schema that is used as a delivery format of data from EDBL. Finally, and before the conclusions, Section 5 illustrates the GUI developed to interact with the database, and Section 6 gives a picture on how we are linking EDBL to other lexical resources in order to enrich it and furnish it with semantic content.

1 The Lexical Database within the Stream of Language Processing Tools

Figure 1 shows the stream of language processing tools developed at the IXA group where the lexical database is integrated. As it can be seen, the lexical information is exported from EDBL into two documents, so distinguishing the general lexicon and the MWLUs' lexicon.
Both lexicons are then used as source at the morphological segmentation process. Later on, the MWLUs' lexicon is taken again as input by the processor of MWLUs.

2 Conceptual Schema of the Database

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

---

[2] EDBL is publicly available for consultation at http://sipl54.si.ehu.es

[3] The authors belong to the IXA research group, devoted mainly to the development of language-processing tools and applications for Basque (http://ixa.si.ehu.es).

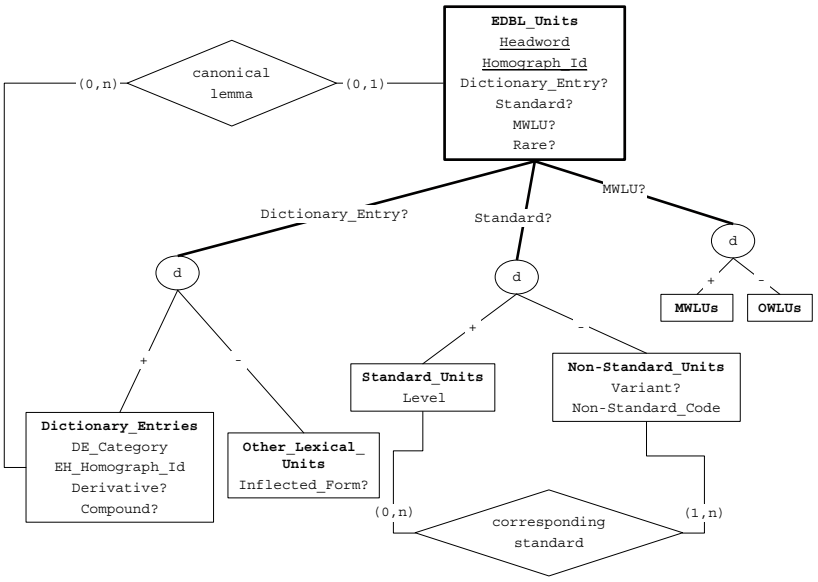

Figure 2: EDBL_Units and the three main specializations.

2.1 Main Entities in EDBL

The main entity in EDBL is EDBL_Units, the key of which is composed of a headword and a homograph identifier, as in any conventional dictionary. The homograph identifier lets us distinguish between entries having the same written form. Every lexical unit in EDBL belongs to this data class. The units in it can be viewed from three different standpoints, and this gives us three different specializations based on the values the units have for certain flag attributes described below (see Fig. 2). All these three specializations are total (meaning that all units in EDBL belong to the three specializations, thick lines in the diagram), and each one establishes a disjointed classification (d within a circle in the diagram) of the units. This classifies every unit in EDBL into (1) standard or non-standard, (2) dictionary entry or other, and (3) one-word or multiword lexical unit.

The canonical lemma relationship between EDBL_Units and Dictionary_Entries links some[4] units in the database such as inflected forms, dialectal variants, etc. to their corresponding lemmas in the Dictionary_ Entries class.

Let us now have a glance at the three main specializations in the following subsections.

2.2 Standard and Non-Standard Lexical Units

Taking into account that Basque is a language still in course of standardization, processes such as spell checking, non-standard language analysis, etc. require information about non-standard entries and their standard counterparts that must be stored in the lexical database.

This specialization divides all the lexical units in EDBL into standard and non-standard entries (see Fig. 3). 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 standard lexical entry. The number of non-standard forms nowadays used in written Basque is still quite large.

The entries belonging to the Non-Standard_Units class can be either variant (mainly dialectal) forms (both at a lexical and at a morphemic level), or simply non-accepted entries.

[4] The cardinality of the relationship is 0:n, meaning that an EDBL unit can be linked via this relationship to several (n) dictionary entries or not linked at all (0). Obviously, most of EDBL units are not related via this relation since their corresponding lemmas are theirselves.

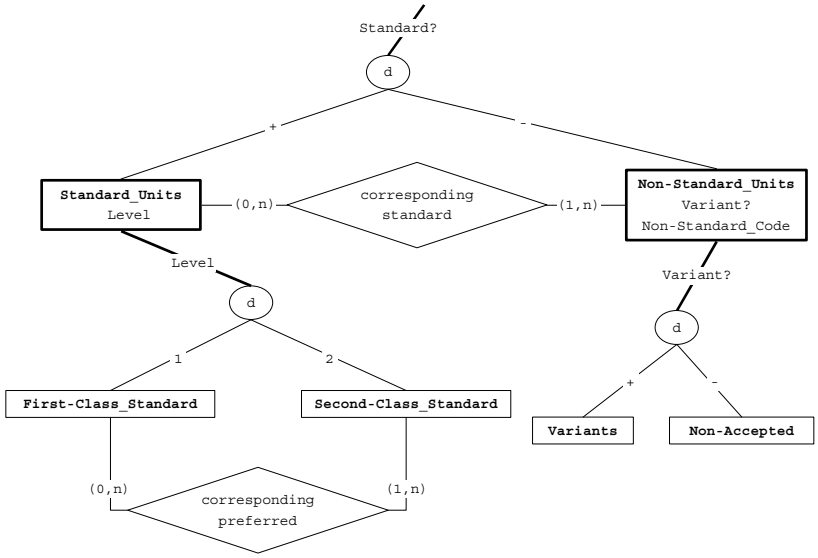

Figure 3: Standard and non-standard lexical units.

The relationship between standard and non-standard units allows us to relate the correct forms to the ones considered incorrect. The existence of more than one dialectal form for any standard entry implies that the participation of `Standard_Units` in this relationship is 0:n, i.e. zero or more non-standard entries can be linked to a standard one. On the other side, each non-standard unit must be related at least to one standard unit (1:n is the participation of non-standard units in the relationship).

Moreover, a disjointed total specialization under `Standard_Units` classifies them into and entries. Second-class standard units are those whose writing is accepted by the academy, but for which it suggests the use of a form. This is the rationale under the relationship linking the `Second-Class_Standard` units to their corresponding first-class entry (the preferred form).

2.3 Dictionary Entries and Other Lexical Units

Another main specialization in EDBL is the one that separates `Dictionary_Entries` from `Other_Lexical_Units` (see Fig. 4).

In the class of dictionary entries, we include any lexical entry that could be found in an ordinary dictionary, and they are further subdivided into nouns, verbs, adjectives, etc. according to their part-of-speech (`DE_Category`). Another specialization divides them into referential entries (symbols, acronyms, and abbreviations), compounds and derivatives.

On the other hand, `Other_Lexical_Units` is totally specialized into two disjoint subclasses: `Inflected_Forms` and `Non-Independent_Morphemes`.

Non-independent morphemes are affixes in general, which require to be attached to a lemma for their use inside a word form.

Some mostly irregular inflected forms are stored in the database within the `Inflected_Forms` class. We have considered here those forms that would need a complex morphotactic treatment (e.g. inflected verb forms) or those that can not be morphologically decomposed in a regular way. The `canonical lemma` relationship explained above links every inflected form to its unique corresponding lemma entry.

2.4 One-Word and Multiword Lexical Units

In order to finish this brief description of the database conceptual schema, we will explain the third total specialization of the main class, which classifies all the units in EDBL into One-Word Lexical Units (OWLUs) and Multiword Lexical Units (MWLUs).

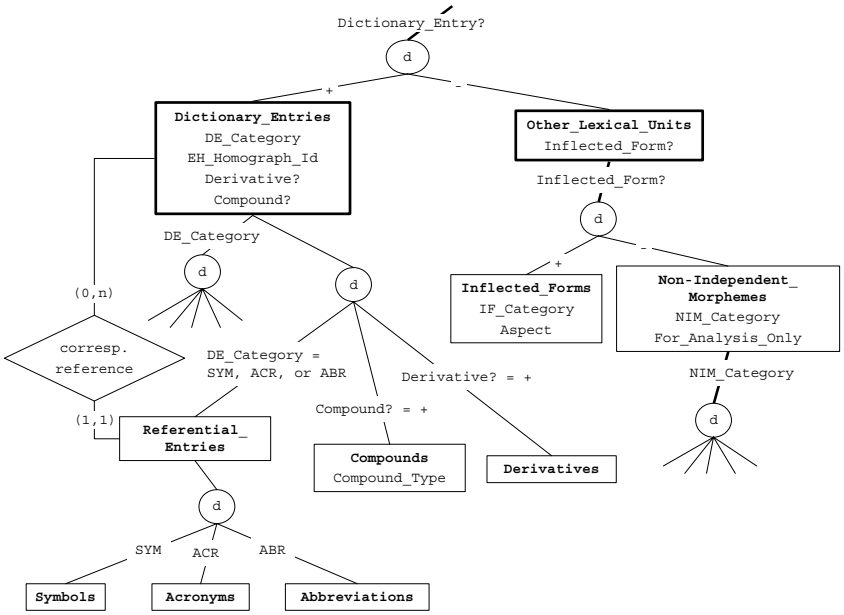

Figure 4: Dictionary entries and other lexical units.

We consider an entry as OWLU if it has not any blanks in its spelling (hyphened forms included). On the other hand, we consider MWLUs all lexical units containing blanks in its spelling.

2.4.1 Morphotactics of One-Word Lexical Units

Every OWLU in EDBL is characterized by its morphotactics, i.e. the description of how it may be linked to other morphemes in order to constitute a word form.

Being an agglutinative language, Basque presents a relatively high power to generate inflected word forms. Any entry independently takes each of the necessary elements (the affixes corresponding to the determiner, number and declension features) for the different functions (syntactic case included). Moreover, noun ellipsis can occur inside a complex noun due to recursive constructions.

We use Koskenniemi's (1983) two-level morphology to represent the morphotactics of Basque word forms, since it is, in our opinion, the most adequate formalism for describing the morphology of agglutinative languages. As it distinguishes the surface and the lexical level of each morpheme, two-level forms of each OWLU are stored in EDBL.

The `morphotactics` relationship (see Fig. 5) is used to describe the different morphological aspects OWLUs adopt. The different entities that take part in this relationship are `OWLUs`, `Two-Level_Forms`, `Continuation_Classes` and `Lexicons`.

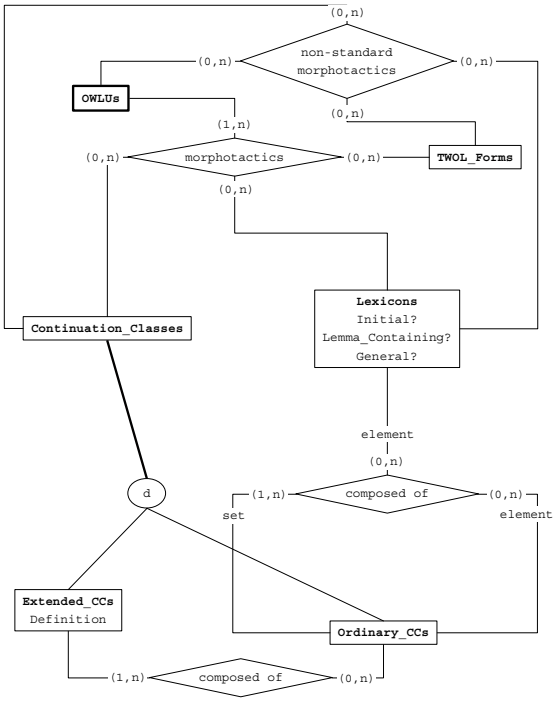

Figure 5: Morphotactics of OWLUs.

Every entry is at least related to one two-level form (two-level forms usually have diacritics by

which the proper morpho-phonological rules will be selected). Every two-level form is related to a lexicon (via the `morphotactics` or the `non-standard morphotactics` relation), as well as included in a continuation class. Relating two-level forms to continuation classes makes defining the set of morphemes that can be added to the stem of an entry possible.

Our lexical system consists currently of approximately 79,000 OWLUs, grouped into 200 two-level sublexicons, and a set of 24 morpho-phonological rules[5] that describe the changes occurring between the lexical and the surface level.

MWLUs' "morphotactical" behavior is described by means of surface realization patterns as described in the next subsection.

2.4.2 Composition and surface realization of Multiword Lexical Units

The description of multiword lexical units within a lexical database must include, at least, two aspects (see Fig. 6): (1) their composition, i.e. which the components are, whether they can be inflected or not, and according to which OWLU they inflect; and (2), what we call the surface realization, that is, the order in which the components may occur in the text, the components' mandatory or optional contiguousness, and the inflection restrictions applicable to each one of the components.

The `composed of` relationship links every MWLU to its components, that may be, in some cases, linked to their corresponding OWLU.

In that what concerns the surface realization, it is to be said that components of MWLUs can appear in the text one after another or dispersed; the order of components is not fixed, as some MWLUs must be composed in a restricted order while others may not: an MWLU's component may appear in different positions in the text; and, finally, the components may either be inflected or occur always in 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 inflection attributes. Moreover, some MWLUs are "sure" and some are ambiguous, since it can not be certainly assured that the same sequence of words in a text corresponds undoubtedly to a multiword entry in any context. According to these features, we use a formal description where different realization patterns may be defined for each MWLU.

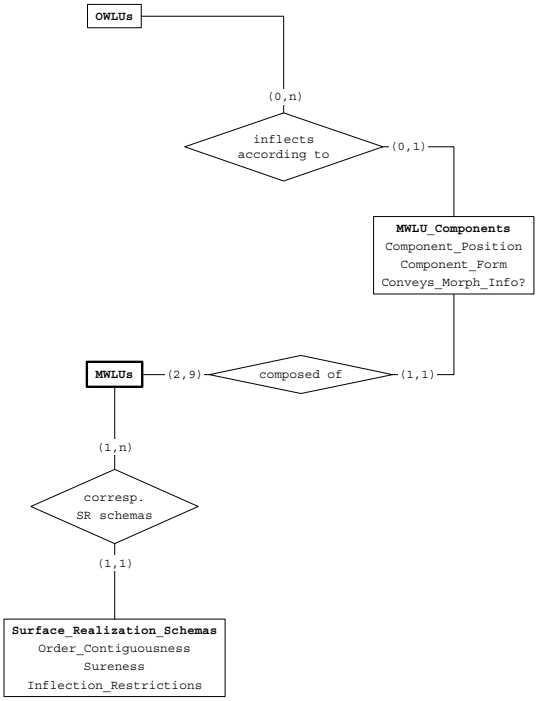

Figure 6: Composition and surface realization of MWLUs.

3 Linguistic Contents

In the previous section, we have seen how all the lexical entries are organized through the database. We will give now some figures to give an idea on the linguistic contents actually stored in EDBL.

According to the classification into the three main specializations, EDBL contains: 59,710 dictionary entries and 20,935 other lexical units (20,588 inflected forms and 347 non-independent morphemes); 77,663 standard forms and 2,982 non-standard; 79,224 OWLUs and 1,421 MWLUs.

All dictionary entries are subclassified into lexical categories that are specified by its corresponding subcategory and any other interesting syntactic or semantic feature. Inflected forms are split up into verbal forms (auxiliary and synthetic verbs) and other

[5] Two-level rules describing morpho-phonological changes are not stored in the database.

inflected forms. Morphosyntactic features such as number, person, mode and aspect are specified for inflected verb forms, and number, case, function and aspect for other inflected forms. Finally, non-independent morphemes are subdivided into different categories (graduators, declension morphemes, etc.) that are specified by required morphosyntactic features.

To the characterization of such classification, a category system has been defined, based mostly on grammars published by the Basque Language Academy.

Among dictionary entries, we distinguish 38,989 nouns (common, person and place names, and figures), 9,664 adjectives (common and interrogative), 6,726 verbs (simple, composed, and periphrastic), 3,441 adverbs (common and interrogative), 60 pronouns (personal, indefinite, reflexive, and reciprocal), 225 determiners (demonstrative, modal, and numeral), 146 connectors (sentence connectors and conjunctions), 124 interjections, 10 particles, 305 referential entries (acronyms, abbreviations, and symbols), and 20 other entries.

Among non-independent morphemes, we have 193 declension morphemes, 44 subordinating morphemes, 37 lexical suffixes, 2 lexical prefixes, 21 graduators, 20 verb type morphemes, 22 aspectual morphemes, and 8 other non-independent morphemes.

4 Mapping the Relational Database into TEI-conformant Feature Structures

In the IXA group, we have designed and almost implemented a plan to integrate and standardize the language processing tools mentioned in Section 1, in such a way that a common data exchange format is used as an input and delivery format between them. This data exchange format is based on SGML-coded TEI-conformant feature structures. Taking into account that EDBL is the first unit in the chain of language processing tools, the information in the database is exported and delivered from it as a collection of feature structures to be used first by the morphological segmentizer.

In order to represent the structure of each one of the entities, typed feature structures (FS) are used. As Ide (1993) point out, 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.

Instead of defining our own formalism for FSs, 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 (FSDs) have been made for all the FS types that are used when exporting data from EDBL. These FSDs 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 `Base-Type` attribute is used in the definition of a type to declare the superclass or basic type from which the type defined inherits features).

So, the conceptual schema of the relational database has been mapped into a hierarchy of typed feature structures. The leaves of this hierarchy are 22 disjoint classes (thick-border boxes in Fig. 7), and each one of them defines a different FS type. The main class of the hierarchy defines the most general structure —`EDBL-Unit-FS`—, whose features are inherited by every class and every instance in the database. Let us show a partial version of the FSD of this main class:

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

In the declaration above, `&Key-Type` and `&Category` are SGML entities that are defined elsewhere. The features in the FSDs may contain different types of values specified by the element `vrange`.

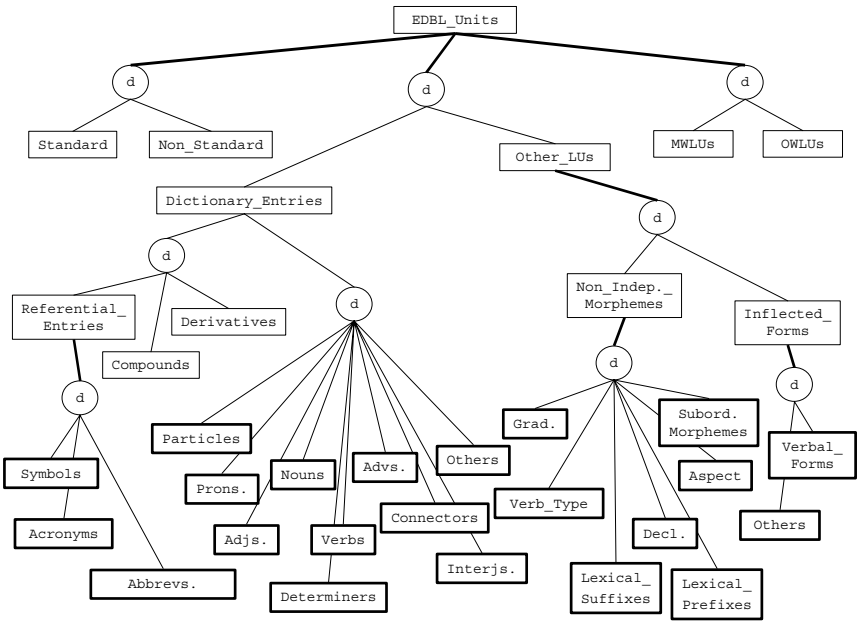

Figure 7: Feature structures' type hierarchy, used to export the information from EDBL.

Let us now show the `Key` FS type, defined by 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="Homograph-Id">
    <fdescr>Homograph identifier used to
            distinguish two units with the
            same headword
    </fdescr>
    <vrange>
      <nbr value="0" rel="ge" type="int">
    </vrange>
  </fdecl>
</fsdecl>
```

When data are exported from EDBL, every EDBL unit is delivered into one of the 22 terminal FS types, including inherited features and others coming from nodes outside the main hierarchy. Let us now show an example of an instance of EDBL, such as it will be delivered. It is the case of a noun, , Basque derivative for . 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 derivative, 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). This is the case of a standard lexical entry (and so it does not convey any data corresponding to non-standard entries) and it belongs to the OWLU's class (information on morphotactics is included).

```
<p>
  <!-- berr+erabil+garri+tasun,
       lexicalized derived noun entry -->
  <fs type="Nouns">
    <f name="Key">
      <fs type="Key-Type">
        <f name="Headword">
          <str>berrerabilgarritasun</str></f>
        <f name="Homograph-Id">
          <nbr value="1"></f>
      </fs>
    </f>
    <f name="Category"><sym value="NOUN"></f>
    <f name="Noun-Features">
      <fs type="Noun-Features-FS">
        <f name="Subcategorization">
          <sym value="Common"></f>
        <f name="Animate"><minus></f>
        <f name="Countable"><minus></f>
        <f name="Measurable"><plus></f>
        <f name="Plural"><minus></f>
      </fs>
    </f>
    <f name="Morphotactics" org="set">
```

```
      <fs type="Morphotactics-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="Derivation-Features">
      <fs type="Derivation-Features-FS">
        <f name="Base">
          <fs type="Key-Type">
            <f name="Headword">
              <str>erabili</str>
            </f>  <!--"to use" verb -->
            <f name="Homograph-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="Homograph-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" suffix-->
            <f name="Homograph-Id">
              <nbr value="1"></f>
          </fs>
          <fs type="Key-Type">
            <f name="Headword">
              <str>tasun</str>
            </f><!-- "-ity" suffix -->
            <f name="Homograph-Id">
              <nbr value="1"></f>
          </fs>
        </f>
      </fs>
    </f>
  </fs>
</p>
```

5 A Browser-Based Graphical User Interface for Lexicographers and Common Users

In order to take advantage of all the information stored, our database has to be accessible and manageable. Even more, the fact that the users will not be computer scientists, but mainly linguists, stresses the reasons why we need a user-friendly, readily accessible and flexible interface.

For that purpose, we designed an Oracle Developer-based interface with these main characteristics:

1. Graphic interface.
2. It provides immediate help to the user based on context.
3. Changeable menus depending on the context.
4. Positive and significant messages and alerts.
5. Flexible interface appearance depending on the user.
6. Accessible from the Internet.

Therefore, we have specially worked on these points:

1. Two levels of access to the database: one that lets users only consult the data, and the second one that gives them full access: query and update.
2. Maintenance tables: for each table in the database, we maintain another table, where all the values an attribute can have are explicitly stated. So, the interface is able to show the users the set of values they can introduce in each field.

6 Linking EDBL to other Lexical Resources

Continuous updating of EDBL is an arduous but necessary task. Moreover, the introduction of semantic content in the database requires linking it to other lexical resources such as machine-readable monolingual dictionaries containing definitions, multilingual dictionaries giving equivalents in other languages, etc.

The entries updating method consists in trying to match the entries in these resources to EDBL entries, and finding which are not yet in the database. For this, we use mainly (Euskaltzaindia, 2000), that is a regularly updated word inventory where standard forms are listed, and non-standard ones are always linked to their corresponding standard. Once the gaps in EDBL are detected, a lexicographer has to decide whether the entries are to be added as standard or non-standard, or not at all.

Other three dictionaries have been used for the same purpose: (Elhuyar, 1998), a Basque-Spanish/Spanish-Basque bilingual dictionary,