# Different issues in the design of a general-purpose Lexical Database for Basque

*Agirre E., Arregi X., Arriola J.M., Artola X.,*
*Díaz de Ilarraza A., Insausti J.M., Sarasola K.*[1]

Department of Computer Languages and Systems
(University of the Basque Country)
E-mail: jiparzux@si.ehu.es

***Abstract***

*EDBL is a lexical database (LDB) for Basque. This paper presents the design and the main features of this database, conceived as a general lexical basis for the automatic treatment of Basque. The conceptual schema of EDBL is explained by means of Extended ER diagrams and Feature Structures. The implementation of the database in a commercial RDBMS and the problems encountered in this implementation are discussed.*

*EDBL, seen as a large repository of lexical information, acts as the basis for a number of different tasks in automatic processing. The applications of the database are presently, and in the short and midterm will be, the following: morphological analysis, spell checking and correction, (semi-)automatic lemmatisation and tagging, syntactic analysis and analysis of textual corpora.*

***Résumé***

[1] E. Agirre and J.M. Arriola are supported in this project by a grant of the Basque Government.

## 1. Introduction.

EDBL (Euskararen Datu-Base Lexikala) was created because of the peremptory need for sound lexical support for the construction of a general morphological analyser and its most important by-product so far, the recently commercialised spelling checker/corrector Xuxen (Agirre *et al.*, 92; Aduriz *et al.*, 93). The maintenance of the large amount of lexical information needed in such a project would not have been possible without a database management system. Other projects our group is currently involved in, like the construction of a lemmatiser/tagger (Aduriz *et al.*, 94) —also derived from the morphological analyser— or the development of a general syntactic parser —based on Constraint Grammar (Karlsson *et al.*, 92)— require different types of lexical information. In this context, EDBL has been redesigned as a general basis for the multiple lexical needs that current and further work on the automatic treatment of Basque will have (we are involved so far in automatic processing tasks of written Basque).

In the beginning, the only application of EDBL was the treatment of morphology. The bias produced because of this first objective of the LDB has been corrected in the new design presented in this paper. The morphological usage of EDBL is not longer its central element, but one more of its several purposes.

The main key of every item in the database is now composed by the headword and an homographe identifier, as in any conventional dictionary. The information is distributed in different parts, according to the different purposes it is intended to be used for.

Another important aspect in this new design of EDBL is that it is conceived both for human users and for natural language applications. A specially designed interface will provide the specialist, that is, the computational lexicographer, with a set of functions that will help them in maintaining and updating the LDB, and in extracting the information needed for the different applications; in the case of the common user, the interface will allow them to use the LDB as if it were an electronic dictionary. Obviously, specifically designed programs get from the LDB all the lexical information that the different NLP applications currently developed require.

After a short section explaining EDBL's most important features, the main part of the paper will be dedicated to presenting the conceptual schema of the database. Finally, after a discussion on the integration of some semantics in the database, a brief description of the mapping of this conceptual structure into the relational model, and a discussion on the problems arisen doing that will be given.

## 2. What is EDBL like?

EDBL's main features are the following:

Multi-purpose. It has been designed as a general support and source for different applications and not for a unique application. Each application gets from EDBL the data it needs. It is currently source of the general lexicon required by the morphological analyser and by the spelling checker, and it will support the

lemmatiser that is being developed. The treatment of Basque syntax will require information that is currently being acquired and recorded in the database. Although all the data are not yet in the LDB, the general structure has been designed with this in mind. Moreover, along with NLP applications, human users have been also taken into account in its new design. In the short term, an explanatory dictionary containing definitions and examples (Sarasola, 84-95) will be integrated into EDBL, thus enhancing it and converting it into an electronic dictionary suitable for human consultation.

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 always open to new descriptions, compatible or not with the previous ones. Actually, the description of the Basque morphology has been based on the well-known two-level model (Koskenniemi, 83) and it is held in the database in this form (diacritics, continuation classes, sublexicons, etc. are used in this description); however, this fact does not constrain the future work, in the sense that if, at sometime, the need for moving to another model is required, the database architecture will allow this conversion with ease.

Open and flexible, so that it will permit, at anytime, adaptation to new goals. It will obviously allow the addition of new information; moreover, as its structure is based on feature structures, it is, in the opinion of the authors, flexible enough to accept new types of information when needed.

User friendly. Conceived for both programs and human users (specialised or not), the interface for the database has to be designed as an easy-to-use tool. This interface is currently being developed.

## 3. Information structure in EDBL: conceptual schema.

### 3.1. Formalism.

The Extended Entity-Relationship (EER) data model is employed to describe the global structure of the database and the relations between the different objects in it. The extended version of the classical ER model is adequate to describe the hierarchical relationships between the entities that are necessary for a proper description of lexical knowledge.

In order to describe the structure of each one of the entities (actually, just some of them will be described), typed Feature Structures (FS) will be used. As is pointed out in (Ide *et al.*, 93), feature structures have been heavily used to encode linguistic information, there exists 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 it is our case, is quite natural.

A Lisp-like notation will be used to show these FS's. Following is the syntax employed in the declaration of the different types of FS's, using an Extended BNF

grammar (EBNF). 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.

EBNF grammar used in the FS type definitions

```
FS_Type_Definition = FS_Type_Identifier "=" [BT] "(" AT {AT} ")".
BT = "(Base-Type :" FS_Identifier ")".  ; base type or superclass
AT = "(" A ":" T ")".
A = Attribute_Identifier.                       ; attribute
T = FS_Type_Identifier | Basic_Type_Identifier. ; type of value
```

### 3.2. Basic entities in the LDB.

The fundamental entity in the LDB is a class (or type of objects) called EDBL Units (EDBL-Unit-FS feature structure type). This class is specialised into three subclasses: Dictionary Entries, that contains those entries in the EDBL that you would expect to find in an ordinary dictionary, Verb Forms, that contains the finite verb forms, and the subclass Non-Independent Morphemes, that mainly contains non-independent morphemes (suffixes, prefixes, etc.).

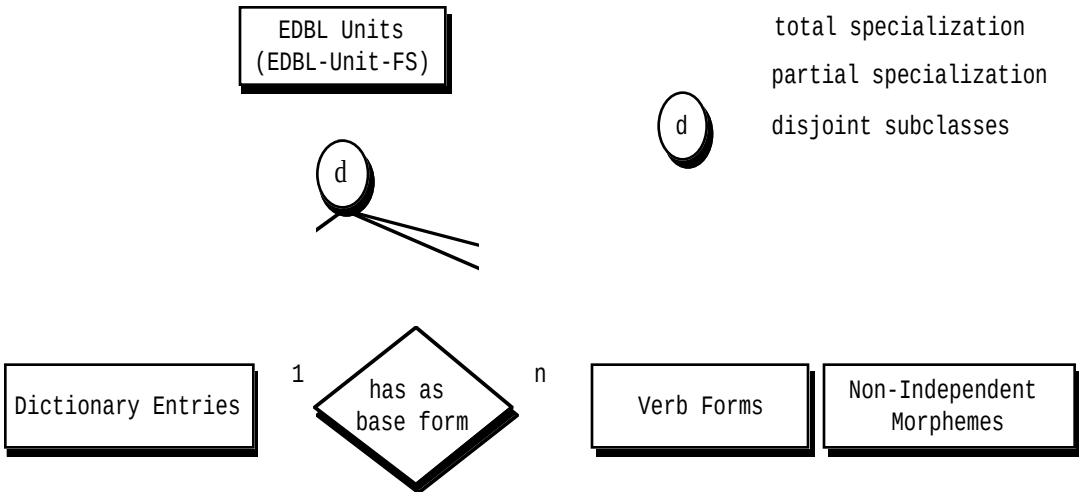


Fig. 1.- General schema of EDBL[2].

In figure 1, the relationships among these four types of objects are shown. The main class has three specialisations —total specialisation (thick line)— resulting into three disjoint subclasses. Between the Dictionary Entries and the Verb Forms

---

[2] As is well known, in this formalism squares represent entities while diamonds stand for relationships. The numbers at both sides of the diamonds indicate the cardinality of the relations. Class-subclass relations between entities are expressed by means of lines linking the entities. The inclusion symbol is usually placed over these lines in order to indicate the sense of the relationship. However, here the convention of understanding the entities placed in lower positions in the diagram as subclasses of the one in upper position (and linked with them by straight lines) is used.

there exists a 1-to-n relationship, that represents which entry is the base or root of each one of the finite verb forms.

Each one of the classes in figure 1 defines a different FS type. The main class defines the most general structure —EDBL-Unit-FS—, inherited by every unit in the database (the semi-colon indicates the beginning of a comment):

```
EDBL-Unit-FS =
     ((Key : Key-FS)
      (POS : POS-Type)                  ; part of speech
      (Morphology : Morphology-FS)
      (Variants : Variant-FS)
      (Source : Source-Type)
      (Source-Form : String))
```

The features in the type definition above may contain different types of values. For instance, the values corresponding to the feature Key must belong to the FS type named Key-FS, composed by two features, Headword and Homographe-Id. This key identifies uniquely every unit in the LDB:

```
Key-FS = ((Headword : String)
          (Homographe-Id : Positive))
```

Other features in the EDBL-Unit-FS definition contain the part of speech of the entry, the morphological information, the variants of the word (dialectal or others), the source dictionary from which the entry was drawn and the exact form the word had in that particular dictionary (the fact that the standardisation of Basque is a process currently still in progress makes for a number of "standard" entries of EDBL differing from its source forms in the particular dictionary).

The two FS type definitions below describe the Verb Forms subclass and the feature structure type defined in it; the attribute Verb-Form will be inherited by every instance of the subclass, that is, by every finite verb form in the LDB:

```
Verb-Forms = ((Base-Type : EDBL-Unit-FS)
              (Verb-Form : Verb-Form-FS))

Verb-Form-FS = ((Base-Form : Key-FS)
                (Mode_Tense : Mode_Tense-Type)
                (Ergative : Ergative-Type)
                (Dative : Dative-Type)
                (Absolute : Absolute-Type)
                (Allocutive : Allocutive-Type))
```

The structure and the different features belonging to the two other main types of objects will be described later, when the morphosyntactic aspects related to their subclasses are discussed.

### 3.3. The morphological component.

As it has been said, the linguistic description of the morphology of Basque is currently based on the well-known computational model called two-level morphology. The information coded in EDBL is used as a source to automatically

extract the lexicons required by the full-coverage morphological analyser and synthesiser MORFEUS (Aduriz *et al.*, 92), a lemmatiser/tagger of unrestricted text that is being currently developed, EUSLEM (Aduriz *et al.*, 94), and the already commercial spelling-checker of Basque Xuxen (Agirre *et al.*, 92).

The morphological aspects of the entries and their variants are described by means of two features that all the lexical units of the database have: Morphology and Variants. The feature called Morphology has as value an FS that contains the two-level form of the word —with diacritics, if necessary, to control the application of the morphophonological two level rules—, and two attributes featuring the morphotactic aspects: the continuation class, that describes the set of morphemes that can follow a given entry word, and the sublexicon to which the entry belongs. The variants of the lexical entry are described also based on the two-level model and are currently employed for a more intelligent correction strategy by the spelling corrector and for the lemmatisation and tagging of non-standard Basque texts. The diagram in figure 2 summarises the entities describing these aspects of the lexical units and their relationships with the other entities around them.

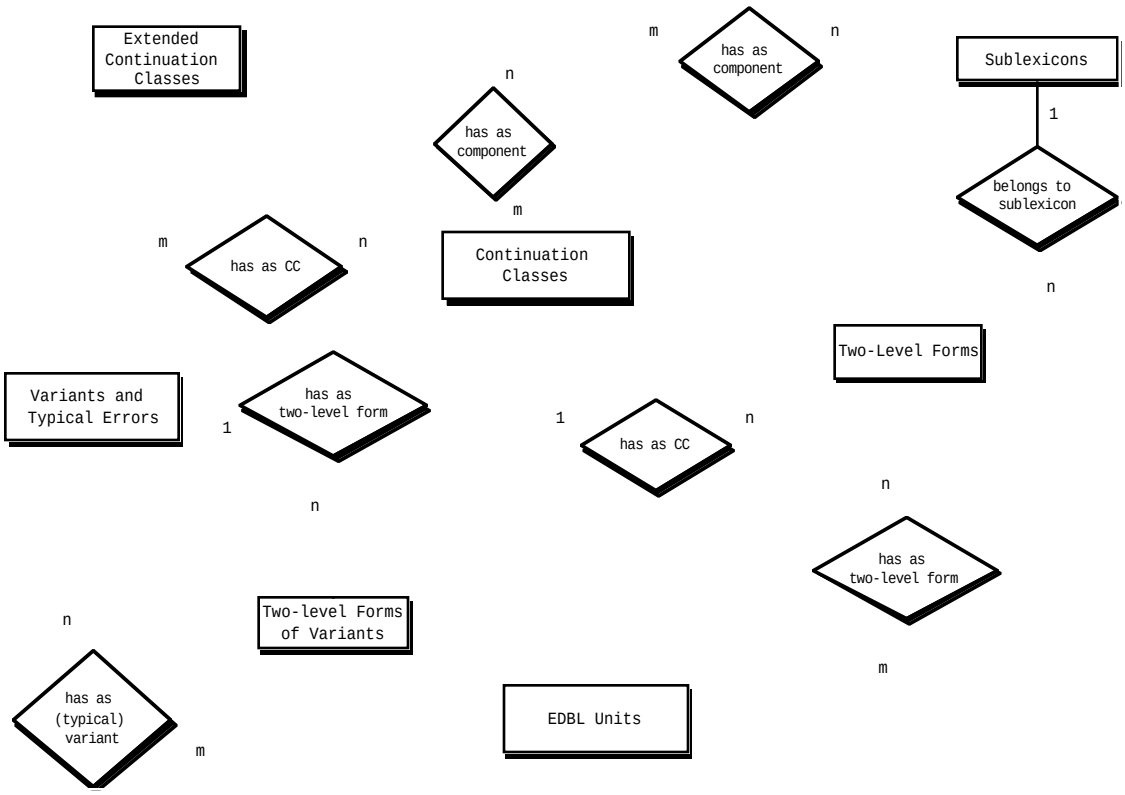


Fig. 2.- Morphological component and variants in EDBL.

The following feature structures are used for the description of the morphophonology and the morphotactics of the entries. Other tables of the database are used to represent the composition of the continuation classes and to describe

the attributes of each one of the sublexicons into which the general lexicon is distributed for morphological tasks.

```
Morphology-FS = ((TWOL-Form : String)
                 (Continuation-Class : Continuation-Class-Type)
                 (Sublexicon : Sublexicon-Type))

Variant-FS = ((Variant-Form : String)
              (Continuation-Class : Continuation-Class-Type)
              (TWOL-Form : String)
              (Error-Code : Error-Code-Type))
```

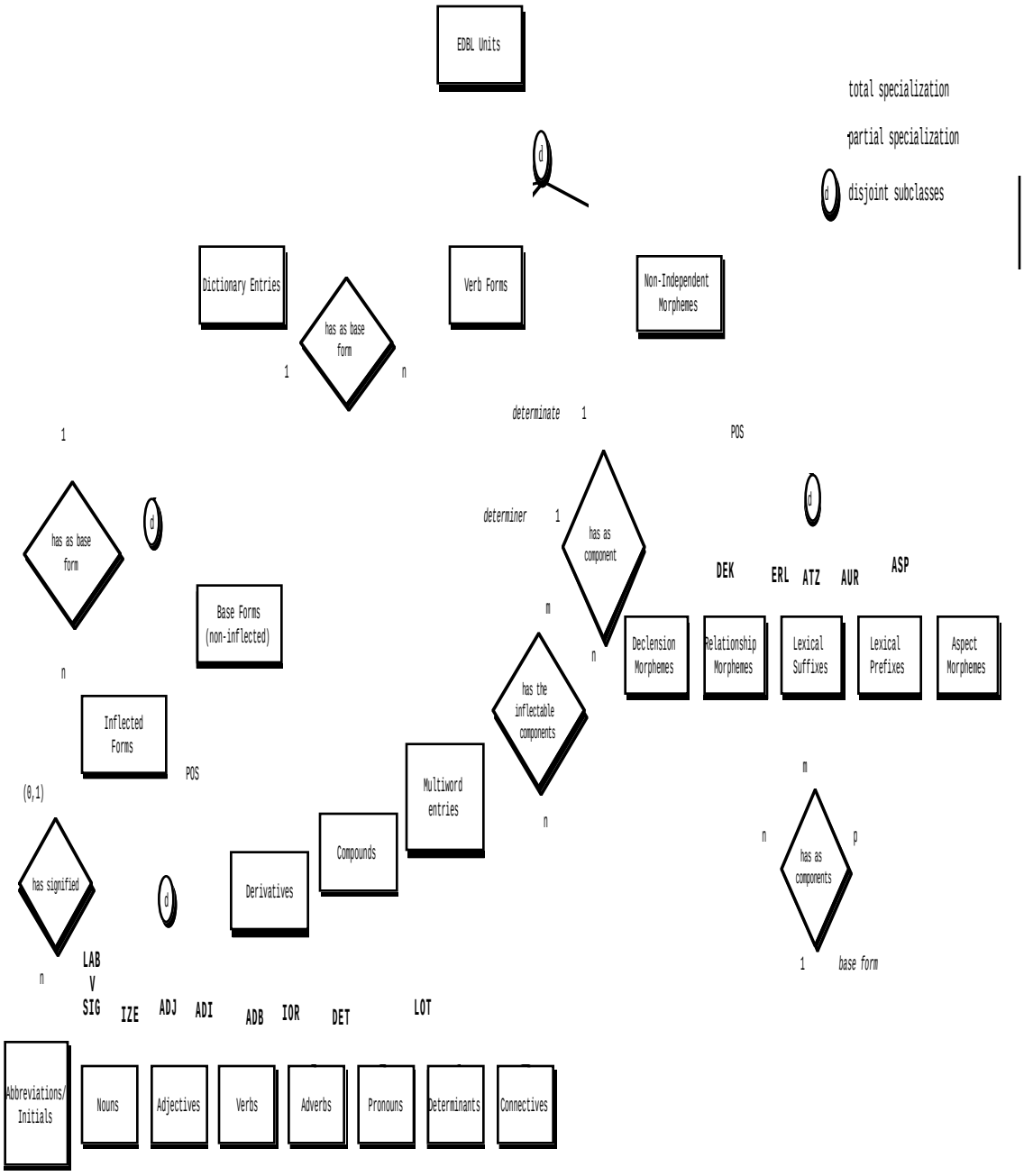


Fig. 3.- Classification of the entries according to morphosyntactic aspects[3].

3 Feature names like POS over the class-subclass relationship lines represent the feature upon whose value the specialisation is realised; for instance, the class of Non-Independent Morphemes is specialised into Declension Morphemes, Relationship Morphemes, etc. according to the values held in the POS feature (DEK, ERL, etc., respectively).

### 3.4. Morphosyntax.

Some of the features describing a dictionary entry or an entry corresponding to a non-independent morpheme are conceived to represent morphosyntactic information. In agglutinative languages like Basque it is not very easy, or, even, it does not make much sense, to separate aspects so tightly related to one each other like morphology and syntax. This is the reason for grouping them together in this description.

The morphosyntactic information of a compound noun, for example, 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 compound, that is, which the determinate of the composition is, which the determiner —both cross-references of other entries in the LDB—, and the type of compound it is. Similarly, a derivative verb entry contains, besides the features describing it as a verb —subcategory, type of auxiliary, and subcategorization pattern—, the features that all the derivatives in the database inherit, that is, the cross-references to its base, and eventually to the prefixes and/or to the suffixes that are present in the word.

The diagram in figure 3 shows the classification of the entries into different subclasses according to the diverse sets of morphosyntactic features their description may have. As has been said, each entry in the LDB will inherit the features from the class it belongs to; in other words, the presence of certain valued features in the description of an entry reveals its belonging to certain class or classes of entries.

In the next section examples of FS's of different types with descriptions of actual entries from EDBL will be explained. The syntax used in these particular instances or classes is described prior to the examples.

### 3.5. Syntax of the instances and actual examples.

Extensions of the model such as those presented in (Ide *et al.*, 93) have been allowed here. These extensions of the basic FS model provide a means of representing the disjunction of several values for a feature, to specify the disjunction of different parts of a FS, etc. The special words LIST and SET are used to denote the two types of disjunctions allowed, that is, the one in which the order of the elements involved is relevant and the other in which it is not:

EBNF grammar used to describe the FS instances

```
FS = "(" T FS_Part {FS_Part} ")".       ; feature structure
T = "( Instance-of " FS_Identifier {FS_Identifier} ")".
                                        ; classes to which the
                                        ; instance belongs
FS_Part = AV | DFS.                     ; part of feature structure
AV = "(" A V ")".                       ; attribute-value pair
DFS = "(" ["LIST"|"SET"] FS ")".        ; general disjunction of
                                          feature structures
V = String | Symbol | FS | VD | "null". ; value
VD = "(" ["LIST"|"SET"]  V {V} ")".     ; value disjunction: list or
                                          set of values
```

Before introducing the actual examples, let us show the definition of several classes of lexical units according to the hierarchy presented in figure 3, and some feature structure types defined for different parts of speech:

Definitions of some lexical unit classes

```
Dictionary-Entries = ((Base-Type : EDBL-Unit-FS))
Base-Forms = ((Base-Type : Dictionary-Entries))
Inflected-Forms = ((Base-Type : Dictionary-Entries)
                   (Inflected-Form : Inflected-Form-FS))
Nouns = ((Base-Type : Base-Forms)
         (Noun : Noun-FS))
Derivatives = ((Base-Type : Base-Forms)
               (Derivation : Derivation-FS))
```

Some feature structure types used in the examples below

```
Noun-FS = ((Subcategorization : Noun-Subcategory-Type)
           (Animate : Animate-Type)
           (Countable : Countable-Type)
           (Measurable : Measurable-Type)
           (Plural : Plural-Type))
Verb-FS = ((Infinitive : String)
           (Subcategorization : Verb-Subcategory-Type)
           (Type-of-Auxiliary : Type-of-Auxiliary-Type)
           (Subcategorization-Pattern :
                                 Subcategorization-Pattern-Type))
Determinant-FS = ((Subcategorization : Determinant-Subcategory-Type)
                  (Number_Definiteness : Number_Definiteness-Type)
                  (Proximity : Proximity-Type)
                  (Position : Position-Type)
                  (Clause-Boundary : Clause-Boundary-Type))
Lexical-Prefix-FS = ((POS-of-Base : POS-Type)
                     (POS-of-Derivative : POS-Type))
Derivation-FS = ((Base : Key-FS)              ; key type value
                 (Prefixes : Key-FS)          ; key type value(s)
                 (Suffixes : Key-FS))         ; key type value(s)
Multiword-Entry-FS =
   ((Discontinuity : Discontinuity-Type)
    (Certainty : Certainty-Type)
    (Order : Order-Type)
    (Inflectable-Constituents : Inflectable-Constituent-FS))
Inflectable-Constituent-FS =
   ((Constituent : Key-FS)                    ; key type value
    (Inflection-Constraints : Inflection-Constraint-Type))
```

Following are given some examples actually extracted from the LDB:

1.- An auxiliary verb finite form, "dio". In Basque, the verb form agrees in person and number with the ergative, dative and absolute cases (the sentence "hura eman dio" is translated into English as "he/she has given that to him/her", being "eman" the Basque word for "to give").

```
((Instance-of Verb-Forms)
 (Key
  ((Headword "dio")
   (Homographe-Id 1)))
 (POS ADL)                ; auxiliary verb
 (Verb-Form
  ((Base-Form
    ((Headword "edun*")   ; "to have": the star means that it is a
                          ; theoretical or reconstructed entry
     (Homographe-Id 1)))
   (Mode_Tense A1)        ; indicative, present
   (Ergative HARK)        ; agreement with the erg. subject: "he/she"
   (Dative HARI)          ; agreement with the dative: "to him/her"
   (Absolute HURA)        ; agreement with the object: "that"
   (Allocutive null)))
 (Morphology
  ((TWOL-Form "dio")
   (Continuation-Class LAT)
   (Sublexicon a125b12378)))
 (Variants null)
 (Source X)
 (Source-Form null))
```

2.- A demonstrative determinant, "hura" (Basque for "that"), with two allomorphs (two two-level forms) and two variants.

```
((Instance-of Determinants)
 (Key
  ((Headword "hura")
   (Homographe-Id 1)))
 (POS DET)                     ; determinant
 (Determinant
  ((Subcategorization ERK)     ; demonstrative
   (Number_Definiteness S)     ; singular
   (Proximity HURA)            ; third degree of proximity
   (Position ATZ)              ; after the noun it determines
   (Clause-Boundary null)))    ; does not necessarily determine
                               ; a clause boundary
 (Morphology                   ; set of values (two morphology FS's)
  (SET ((TWOL-Form "haQ")
        (Continuation-Class E3)
        (Sublexicon lemak))
       ((TWOL-Form "hura")
        (Continuation-Class I0)
        (Sublexicon lemak))))
```

```
 (Variants                      ; set of values (two variants)
  (SET ((Variant-Form "ura")
        (Continuation-Class E3)
        (TWOL-Form "aQ")
        (Error-Code DIAL))
       ((Variant-Form "ura")
        (Continuation-Class I0)
        (TWOL-Form "ura")
        (Error-Code DIAL))))
 (Source K)
 (Source-Form "hura"))
```

In this case, it is interesting to notice that, applying general disjunction to the Morphology FS, the Sublexicon feature can be factored, thus avoiding redundance and permitting a cleaner representation (similarly, Variant-Form and Error-Code could also be factored in the Variants FS):

```
( ...
 (Morphology
  ((Sublexicon lemak)            ; factorisation of Sublexicon
   (SET                          ; general disjunction, within the
                                 ; Morphology FS
    ((TWOL-Form "haQ")
     (Continuation-Class E3))
    ((TWOL-Form "hura")
     (Continuation-Class I0)))))
 ...)
```

3.- A derivative noun, "berrerabilgarritasun", Basque for "reusability".

```
((Instance-of Nouns Derivatives)
 (Key
  ((Headword "berrerabilgarritasun")
   (Homographe-Id 1)))
 (POS IZE)                      ; noun
 (Noun
  ((Subcategorization ARR)      ; common noun
   (Animate -)
   (Countable -)
   (Measurable +)
   (Plural -)))
 (Derivation
  ((Base
    ((Headword "erabili")       ; "to use"
     (Homographe-Id: 1)))
   (Prefixes
    ((Headword "ber")           ; "re-"
     (Homographe-Id 1)))
   (Suffixes                    ; list of values: order is relevant
    (LIST ((Headword "garri")   ; English "-ble"
           (Homographe-Id 1))
          ((Headword "tasun")   ; English "-ty"
           (Homographe-Id 1))))))
 (Morphology
  ((TWOL-Form "berrerabilgarritasun")
   (Continuation-Class I)
   (Sublexicon izenak)))
 (Variants null)
```

```
(Source IH)
(Source-Form "berrerabilgarritasun"))
```

This entry belongs to the LDB because it can be said that it is already lexicalised in the common use of the language. However, the description of the different components of this lexical entry could be made in such a way that, should it were not an entry as such, it would still be recognised.

4.- A compound noun, "sistema eragile", Basque for "operating system", multi-word entry.

```
((Instance-of Nouns Compounds Multiword-Entries)
 (Key
  ((Headword "sistema eragile")
   (Homographe-Id 1)))
 (POS IZE)                          ; noun
 (Noun
  ((Subcategorization ARR)          ; common noun
   (Animate -)
   (Countable +)
   (Measurable null)
   (Plural -)))
 (Composition
  ((Determinate
    ((Headword "sistema")           ; "system"
     (Homographe-Id: 1)))
   (Determiner
    ((Headword "eragile")           ; "operating"
     (Homographe-Id 1)))
   (Type-of-Compound I+ADJ)))       ; noun + adjective
 (Multiword-Entry
  ((Discontinuity -)
   (Certainty -)
   (Order +)
   (Inflectable-Constituents        ; only one component is susceptible
                                    ; of inflection in this case
    ((Constituent
      ((Headword "eragile")
       (Homographe-Id 1)))
     (Inflection-Constraints (NOT GRAD))))))
                                    ; the inflection of graduation of
                                    ; the adjective is not allowed
 (Morphology
  ((TWOL-Form "sistema_eragile")
   (Continuation-Class I)
   (Sublexicon izenak)))
 (Variants null)
 (Source IH)
 (Source-Form "sistema eragile"))
```

The description of entries like this is contemplated in EDBL. Although not totally operative in the actual versions of MORFEUS and EUSLEM, the analysis of this kind of multi-word entries (locutions, idiomatic phrases, multi-word terms, etc.) is currently being faced and some results are expected shortly (Aduriz *et al.*, 94).

5.- A currently productive lexical prefix, "ber", that prepended to a verb produces a derivative verb (equivalent to English "re-" in "reheat").

```
((Instance-of Lexical-Prefixes)
 (Key
  ((Headword "ber")
   (Homographe-Id 1)))
 (POS AUR)                          ; lexical prefix
 (Lexical-Prefix
   ((POS-of-Base ADI)               ; verb
    (POS-of-Derivative ADI)))       ; verb
 (Morphology
  ((TWOL-Form "beR")
   (Continuation-Class ADITZAK)
   (Sublexicon aurrizkiak)))
 (Variants null)
 (Source X)
 (Source-Form null))
```

## 4. Semantics in EDBL.

As has been said, this LDB was originally created as a tool for the morphological analysis of Basque, and now we are adding features of a more "purely syntactic" nature, as a result of our current work on computational syntax. The treatment of sentence semantics is not yet one of our most urgent tasks, but we aim to include some lexical semantics in the database in the near future.

The design of EDBL is open and flexible enough to accept the necessary kind of information, due to its use of feature structures.

The inclusion of an homographe identification in the key that distinguishes uniquely all the entries in EDBL demonstrates its "semantic vocation". Certain features already defined in the different classes of lexical units also show that this aspect has been important in the design of the database. Moreover, the classification of the different types of units into subclasses —mainly into the different parts of speech— will ease the addition of semantic features in a specialised way.

Work currently being developed on semantics-based correction in our group has revealed the kind of semantic information required to deal with the concrete problem of selecting, from a set of proposals given by the spelling corrector, the correct one (Agirre *et al.*, 94a): selectional restrictions, for example.

Shortly, an explanatory dictionary containing definitions and examples will be integrated into EDBL; this will be done automatically from a machine readable version of the dictionary. The first consequence of this is that polysemy will be introduced. Besides that, from the analysis of the definition sentences of this dictionary, certain lexical-semantic information will be extracted and integrated into EDBL: taxonomic relationships, synonymy, meronymic relations, and so on. The integration of all this information will become a good testing bench for the design of the database.

## 5. Current implementation of EDBL and some problems.

The conceptual schema of EDBL has been mapped into the relational data model, and the database is physically stored in a commercial RDBMS (ORACLE). It contains currently around 60,000 entries.

The primary key —headword plus homographe identification—, which is often the object of cross-references, is converted in the implementation into a sequence number, in order to avoid costly updates, minimise errors, and provide a greater warranty of consistency of the data. Eventual modifications on the spelling of the headword are only allowed in the tables corresponding to the main types of units.

The different tables of the database have been grouped into several sections, according to the conceptual schema outlined above: main tables (corresponding to the three main types of lexical units), tables related to the morphology of the words, tables that describe variants and common misspellings, tables containing morphosyntactic information —specialised for each one of the different subclasses into which the main classes have been divided—, tables intended to contain semantic information and others (database maintenance, etc.).

This mapping has not been done without problems. The impossibility of operations like factorisation, for example, has led us to need redundant information in some cases. Multi-valued features (sets or lists) have had to be implemented in not very elegant ways, limiting, for example, the number of values that the feature is allowed to have, etc.

The introduction of semantics shortly to be tackled indicates some of the problems that will undoubtedly arise: for example, the subcategorization pattern of some verbs corresponds to the whole homographe —to all its senses— whereas in other cases different patterns are to be taken into account for each one of the senses of the verb. This will inevitably lead us, at least while we keep mapping feature structures into the relational model, to the need of repeating information for the different senses of a word in many cases.

The solution proposed in (Ide *et al.*, 93) consists of mapping the feature structures into the object oriented model. Our next step will be the study of different OODBMS's in order to examine their adequacy to solve, in a more elegant and efficient way, the representation problems encountered so far.

## 6. Conclusion.

In this paper the design of a general-purpose lexical database has been described. The database is mainly used as a source and support for the automatic treatment of written Basque. It contains not only standard dictionary entries but also dialectal variants, finite verb forms and some other inflected forms, non-independent morphemes, compounds, multi-word entries, abbreviations, etc.

Extended Entity-Relationship diagrams and Feature Structures have been employed to describe the conceptual schema of the database. These models have been shown as suitable for the description of the lexical entities, their attributes, and their inter-relationships.

The implementation of the system as a relational database and, thus, the facilities that a commercial RDBMS provides with, make EDBL a practical and very useful tool that is actually used in all that is related with the automatic treatment of the language. However, the mapping of the conceptual schema into the relational model is not completely satisfactory. New ways to do it, mainly in the application of the object oriented model as a more suitable data model for lexical knowledge, are being studied.

## 7. Acknowledgement.

Prof. Jon Patrick (Information Systems. Massey University, New Zealand) for his fruitful comments on an earlier version of this paper.

## 8. References.

Aduriz I., Agirre E., Alegria I., Arregi X., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., Maritxalar M., Sarasola K., Urkia M.. "A morphological analyzer for Basque based on two-level morphology" , Proceedings of the 5th. International Morphology Meeting. Krems (Austria), 1992.

Aduriz I., Agirre E., Alegria I., Arregi X., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., Maritxalar M., Sarasola K., Urkia M.. "Morphological Analysis Based Method for Spelling Correction" (poster session), Proceedings of the European Association for Computacional Linguistics, EACL'93 (Utrecht, The Netherlands), p. 463. 1993.

Aduriz I., Aldezabal I., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Urkia M.. EUSLEM: un lematizador/etiquetador de textos en euskara, Actas del X Congreso SEPLN (Córdoba, Spain). 1994.

Agirre E., Alegria I., Arregi X., Artola X., Díaz de Ilarraza A., Maritxalar M., Sarasola K., Urkia M.. "XUXEN: A spelling checker/corrector for Basque based on two-level morphology", Proceedings of the 3rd. Conf. on Applied Natural Language Processing (ANLP'92, Trento), pp. 119-125, 1992.

Agirre E., Arregi X., Artola X., Díaz de Ilarraza A., Sarasola, K.. Conceptual Distance and Automatic Spelling Correction, Proc. of the Workshop on Comp. Linguistics for Speech and Handwriting Recognition (forthcoming). Leeds (Great Britain), 1994a.

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

*An Overview of the EDR Electronic Dictionaries*. Japan Electronic Dictionary Research Institute, Ltd. TR-024, 1990.

*CELEX News* (newsletters 1 to 5). Centre for Lexical Information, University of Nijmegen (The Nederlands). 1986-1990.

*EDR Electronic Dictionary Technical Guide*. Japan Electronic Dictionary Research Institute, Ltd. TR-042, 1993.

Euskaltzaindia. *Aditz laguntzaile batua*. Euskaltzaindia: Bilbo. 1973.

Euskaltzaindia. *Euskal Gramatika: Lehen urratsak (I eta II)*. Euskaltzaindia: Bilbo, 1985.

Ide N., Le Maître J., Véronis J.. Outline of a Model for Lexical Databases. *Information Processing and Management*, vol. 29, no. 2, pp. 159-186, 1993.

Ingria R.. Lexical Information for Parsing Systems: Points of Convergence and Divergence,*Workshop "Automating the Lexicon"* (Grosseto). 1986.

Karlsson F., Voutilainen A., Heikkilä J., Anttila A. eds. *Constraint Grammar: A Language-independent System for Parsing Unrestricted Text.* Manuscript. Dept. of General Linguistics, Univ. of Helsinki. 1992.

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

Pin-Ngern Colon S.. *A Lexical Database for English to Support Information Retrieval, Parsing, and Text Generation.* PHD Thesis, Illinois Institute of Technology. Chicago, 1990.

Pin-Ngern Colon S., Evens M., Ahlswede T., Strutz R.. Developing a Large Lexical Database for Information Retrieval, Parsing, and Text Generation Systems. *Information Processing and Management*, vol. 29, no. 4, pp. 415-431, 1993.

Sarasola I.. *Hauta-lanerako euskal hiztegia.* Gipuzkoa Donostia Kutxa: Donostia, 1984-1995.

Zampolli A.. Perspectives for an Italian Multifunctional Lexical Database, in A. Zampolli ed., 301-341, *Studies in Honour of Roberto Busa S.J.* Pisa: Giardini, 1987.