A Conceptual Schema for a Basque Lexical-Semantic Framework

ENEKO AGIRRE, OLATZ ANSA, XABIER ARREGI, XABIER ARTOLA, ARANTZA DIAZ DE ILARRAZA, MIKEL LERSUNDI

Computer Science Faculty - University of the Basque Country
P.O. box 649, E-20080 Donostia
Basque Country
e-mail: jipanoso@si.ehu.es

Abstract

This article describes the conceptual schema of a lexical-semantic database for Basque. The schema lets us represent both the Basque version of EuroWordNet and the implicit lexical-semantic information extracted in a semi-automatic way from a Basque monolingual dictionary.

The model presented here has the following features: i) it is general: that is, it subsumes the structure of EuroWordNet, so that we can represent different types of relations between words, senses, and even between synsets, ii) it is suitable for real applications: it has been implemented in a conventional database management system (Oracle 8©), in order to guarantee the practical use of the knowledge base, iii) it is linked to other lexical resources, in order to configure a general lexical framework for Basque language processing.

This general lexical-semantic framework is being supplied with the information extracted from the dictionary, and the construction of the Basque WordNet knowledge base is in an advanced state.

1. Introduction

In this article we introduce the Basque Lexical Knowledge Base (BLKB). It is a large store of lexical-semantic information that has been conceived as a multi-purposed lexical-semantic support, i.e. a goal-independent resource for processing the language. Application-oriented lexicons will be obtained from such a general support.

BLKB is composed of three databases: EDBL, Basque WordNet, and DKB (cf. figure 1). Each of them can be described as follows:

- EDBL (*Euskararen Datu-Base Lexikala*) (Aldezabal et al., 2001) is the lexical database for Basque. It contains grammar information about more than 80,000 entries. The lexicons obtained from it are subsequently used in tools such as a morphological analyser, a spelling checker, a tagger/lemmatiser, etc.
- Basque WordNet (Agirre et al., 2002a) is a lexical-semantic knowledge base whose point of departure is the English WordNet developed at Princeton University (Fellbaum, 1998). Taking the English WordNet as a reference, new wordnets have been built for some other languages, especially in the framework of the EuroWordNet project[1] (Vossen et al., 98, 01). EuroWordNet, basically, adds multilingual links across different WordNets. In this context, Basque WordNet is the Basque component within EuroWordNet.
- DKB (Dictionary Knowledge Base) is being built in a semi-automatic way from a monolingual Basque dictionary (Sarasola I., 1996). It contains a dictionary database as well as lexical-semantic information extracted from the dictionary.

The main core of the article is the description of the DKB database, and it is structured as follows: firstly a general view of BLKB is presented; in section 3 the representation model of the DKB is described; section 4 explains how Basque WordNet can be represented in this schema; in section 5 we present the current state and future work, and finally some conclusions are drawn.

2. General structure of the Basque Lexical Knowledge Base (BLKB)

BLKB is a general framework for Basque lexical-semantic information. This framework is composed by several databases, which are connected in pairs (cf. figure 1) and can be described as follows.

Basque WordNet and EuroWordNet are inherently related, as the first one is a component of the second. The Basque WordNet is being created dynamically by establishing links between Basque senses and EuroWN synsets. This way of proceeding facilitates its construction and allows the storage of multilingual relations.

Basque WordNet is mainly a way of classifying concepts. A similar task has been accomplished by analysing the definitions of a dictionary. The results of such an analysis are stored in the Concept Classification part of the DKB, which can be related to Basque WordNet. It is worth underlining that criteria followed in the creation of both databases are quite different, and so are the obtained relations. Therefore, the integration (total or partial) of these databases allows mutual enrichment.

The DKB groups two different views of the dictionary data. The Dictionary Database stores the dictionary itself, and the Lexical-Semantic Database stores the lexical-semantic relations extracted from it. Within the Lexical-Semantic Database we distinguish between the Concept Classification level, used for storing concepts and relations between them, and the Concept Contextualisation level, where usage instances of these concepts are stored.

[1] http://www.hum.uva.nl/~ewn

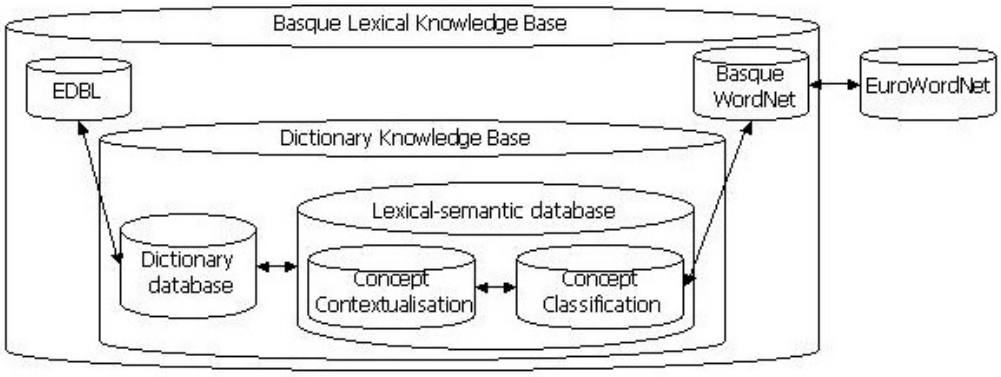

Figure 1: General structure of the Basque Lexical Knowledge Base (BLKB)

Finally, EDBL and DKB are related too. Their inter-relation allows us to manage lexical information of both grammatical and semantic nature, given that EDBL stores the grammatical information of words.

3. Representation of the lexical-semantic information in DKB

3.1. Motivation

In the context of the increasing importance of lexicons in Natural Language Processing, we considered the need to build a lexical knowledge base for Basque. Previously, some studies on lexical knowledge representation have been carried out in the IXA group (Agirre et al., 1994). A French knowledge base was built from the *Le Plus Petit Larousse* French monolingual dictionary in a semi-automatic way (Artola X., 1993). The prototype of the system was implemented using KEE (*Knowledge Engineering Environment*). Its implementation was not easily portable to a conventional environment.

Recently, two works have been carried out in order to create lexical-semantic resources for the Basque language: i) the storing of synonymy and hyponymy/hypernymy relations in the multilingual EuroWordNet project, and ii) the extraction of more relations (synomyny, antonymy, cause, manner, theme…) apart from the already mentioned hyponymy/hypernymy in the analysis of the monolingual dictionary (Agirre et al., 2000; Agirre & Lersundi, 2001; Agirre et al., 2002b).

Therefore, we are interested in developing a general lexical-semantic framework, in which all type of relations (even multilingual and complex ones) are incorporated. Moreover, we want this lexical resource to be usable in practical applications.

3.2. Dictionary database

This part of the DKB stores the dictionary itself with definitions and examples of the different senses. So, for every concept in the lexical-semantic database we have its definition and usage example, if it appears in the dictionary.

We are just going to describe the main entities and relations of this database schema. The main entity in the Dictionary database is `Words`, whose key is composed of a headword and a homograph identifier, just as in the dictionary. A disjointed total specialisation under `Words` classifies them into `Entries` and `Subentries`. The relationship placed between `Entries` and `Subentries` allows us to associate any subentry to its entry. The existence of more than one subentry-form for any entry implies that the participation of `Entries` in this relationship is 0:n.

Moreover, the Words have different senses and they are related by the have_senses relationship. Each Word must be related at least to one sense (1:n is the participation of Words in the relationship).

Finally, having Words and Senses many attributes in common, they have been merged in a Dictionary-Units superclass.

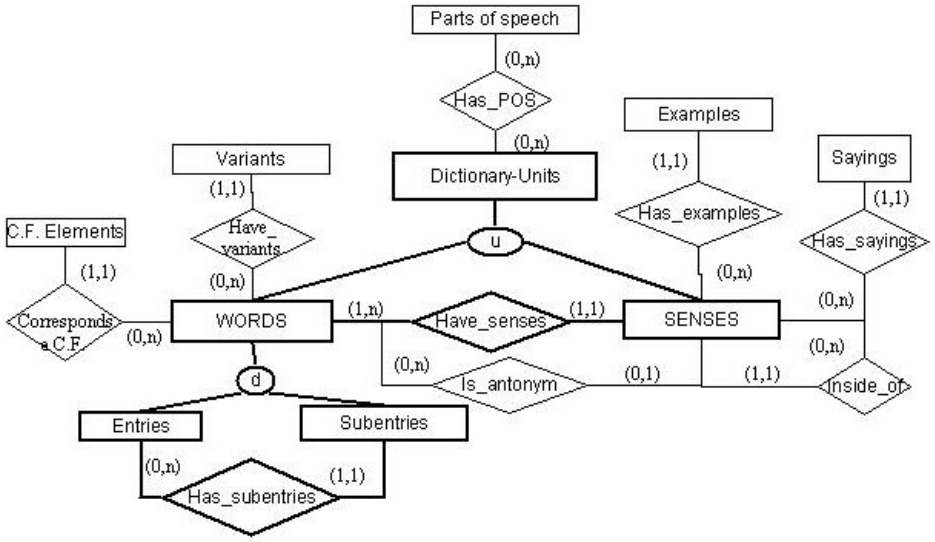

Figure 2: Schema of Dictionary database

3.3. Lexical-semantic database

This database is aimed to store lexical-semantic relations between concepts. Two different types of concepts have been distinguished: *type concepts* and *phrasal concepts*. The basis of this distinction comes from Quillian's (Quillian, 1968) type concepts and tokens.

Type concepts are used to represent context-independent features. The attributes and relations of the type concepts occur for every instance of the concept in every context, so they are intrinsic. An example of this kind of type concepts can be seen in section 3.3.3. The concepts goad and stick are related, and this relation is context-independent.

Phrasal concepts refer to the appearance of a concept in a specific phrase in meaning definitions. The attributes and relations that correspond to the phrasal concepts are related in this concrete context. In section 3.3.3 we illustrate this kind of phrasal concepts by means of an example. The concepts stick and sharp are only related in the definition of goad, but we cannot infer that these two phrasal concepts will have the same relation in any context they appear.

In a similar way, we distinguish two relational levels. On the one hand, we specify the concept-classification level, which stores the type concepts and relations between them. On the other hand, we define the concept-contextualisation level, which stores the phrasal concepts and the relations between them, as well as the relations between the type concepts and their occurrences in specific contexts. That is, if two concepts are related in a definition, then this relation is stored in the concept-contextualisation level, but we cannot always infer that these two concepts are intrinsically related.

So, relations in the concept-classification level represent prototypical uses of senses or words, whereas relations in the concept-contextualisation level represent occurrences.

The information extracted semi-automatically from the dictionary needs to be processed before being stored in the DKB. The information is stored in the DKB, after word-sense disambiguation and relation disambiguation processes are carried out.

3.3.1. Concept Classification

As we mentioned previously, the objective of this work is to store lexical-semantic relations between concepts. In the concept-classification level, we consider as concept any sense of the dictionary, so we could say that the objective is to store lexical-semantic relations between senses.

Being aware that fully disambiguated knowledge base cannot be achieved in a short term, in the meanwhile we decided to store four possible states: sense-to-sense lexical-semantic relations, sense-to-sense syntagmatic relations, sense-to-word lexical-semantic relations and sense-to-word syntagmatic relations. We took this decision because the knowledge base must be the support for later processes and handwork. Furthermore, we consider that even the sense-to-word syntagmatic relations could be useful in some applications.

The four possible states mentioned before are represented in the conceptual schema (cf. figure 3). As we can see in the figure, there are two entities, `Words` and `Senses`, which belong to the `Dictionary database`. There are two more entities `Syntagmatic relations` and

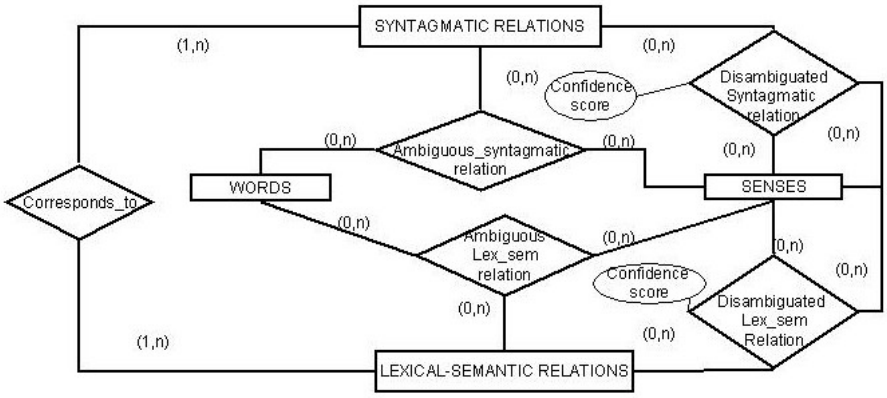

Figure 3: Schema of concept classification level

`Lexical-semantic relations` to store the attributes of the relations we use. The set of relations is defined according to the nature of the syntactic or lexical-semantic information extracted from the dictionary. `Syntagmatic relations` are linked under the `Corresponds_to` relationship to the `Lexical-semantic relations` unit to recognize which lexical-semantic relation can be under the syntagmatic ones. The `Ambiguous_syntagmatic relation`, `Ambiguous_Lex-sem relation`, `Disambiguated_Syntagmatic relation` and `Disambiguated_Lex-sem relation` relationships are used to describe the four possible states detailed above, i.e. sense-to-word syntagmatic relation, sense-to-word lexical-semantic relation, sense-to-sense syntagmatic relation and sense-to-sense lexical-semantic relation, respectively.

3.3.2. Concept Contextualisation

In this part of the representation model the goal is to store relations between phrasal concepts, which are obtained from the dictionary definitions. Each definition is represented by a phrasal concept. The phrasal concepts are defined recursively, that is, within a phrasal concept we can find another one.

These phrasal concepts are specialised into `compound` and `simple` units. This specialisation is total (meaning that all phrasal concepts belong to any of these subclasses) and disjointed (`d` within a circle in the diagram, meaning that a phrasal concept can only belong to one of these subclasses). The head of these phrasal concepts is an instance of one concept of the Concept

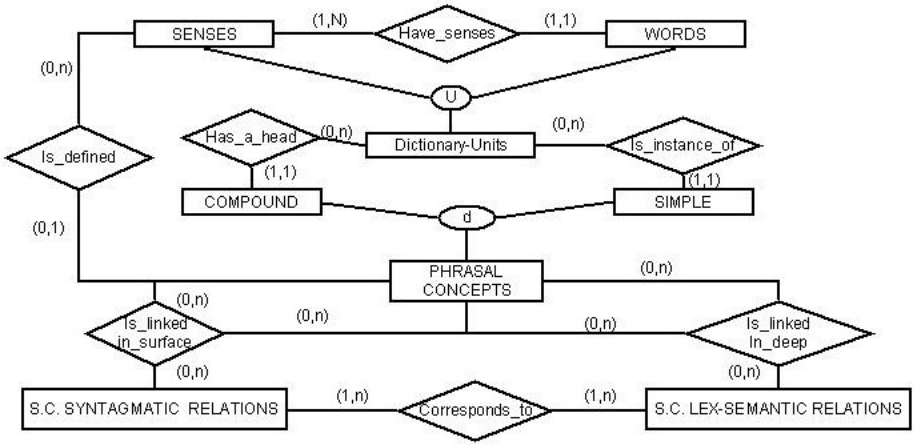

Figure 4: Schema of concept contextualisation level

Classification part. This instance will be related to its concept using `has_a_head` and `has_instance` relationships in the diagram. The instance will be related to the sense if the sense disambiguation process succeeds. If not, it will be related to the word. As we said before (see section 3.2.1), `Words` and `Senses` merge into the superclass `Dictionary-Units`. Therefore, it is not necessary to duplicate `has_a_head` and `has_instance` relationships.

`Is_linked_in_surface` and `is_linked_in_deep` are two relationships that relate two phrasal concepts. If the relation between two phrasal concepts is disambiguated, that is, we obtain the lexical-semantic relation, it will be stored using `is_linked_in_deep` relation. Otherwise, if we only obtain the syntagmatic relation, we store it using `is_linked_in_surface` relation.

3.3.3. Example.

Let us consider the following definition of the dictionary, with its corresponding translation.

Akuilu A1: *Makila luze eta buru-eztenduna, abereei eragiteko erabiltzen dena.*

Goad A1: Sharp, pointed stick that is used for driving cattle.

In figure 5, we can see how the items of concept classification and concept contextualisation levels are related.

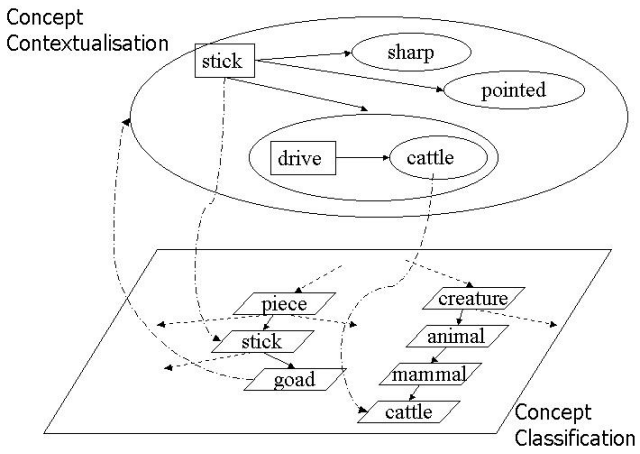

Figure 5: General view of Lexical-Semantic database

At the concept classification level, we see a part of the obtained hierarchy. Each of the concepts in the hierarchy will be related to a phrasal concept belonging to the concept classification level. In the figure, `goad` is related to the phrasal concept obtained from its definition. Finally, we can also see how senses in concept classification (`stick` and `cattle`) are related to their uses in the concept contextualisation level.

4. Representation of the Basque WordNet in this schema

We manage the representation of the Basque version of EuroWordNet by reusing the model of the Concept Classification part described above. Before reasoning on the suitability of the model, we want to remark that, from our point of view, the types of relations defined in WordNet are not enough to describe some interesting linguistic phenomena. In fact, the need of managing a broad set of lexical relations has motivated us to incorporate new types of relations. Therefore, our proposal does subsume the EuroWordNet model, and in addition it incorporates more possibilities.

In table 1, we show the set of types of relations that are dealt with, attending to the basic entities of the enriched Basque WordNet: synsets, variants, and words.

RELATION-TYPES	USABILITY
word - variant	topic signatures (Agirre et al., 2001)
variant - synset	selectional restrictions (Agirre & Martinez, 2001)
synset - synset	selectional restrictions
word - synset	selectional restrictions
word - word	Collocations
variant - variant	Antonyms

Table 1: Set of types of relations

At this point, we must define the categories of WordNet in terms of the entities of our schema.

In our opinion all the relation-types above-mentioned can be correctly represented in our model just by establishing a set of correspondences between the entities of Basque WordNet and the entities of the general representation schema of the Concept Classification. According to that, words in Basque WordNet correspond to the `Words` entity in the schema, variants in Basque WordNet correspond to `Senses`, and synsets in Basque WordNet are represented by the explicit `synonymy` relation in the schema. Following this criterion, the proposed schema is suitable to represent Basque WordNet, including the new relation-types.

5. Current state of the BLKB and future work

The dictionary database in DKB has been already supplied with the information extracted from the dictionary. Namely, 33,102 dictionary-units, 3,160 sub-entries, and 45,873 senses with their corresponding relations are stored in the Dictionary Database.

Besides, the links between the DKB and the lexical database (EDBL) have been established. Table 2 shows the level of integration between the two databases.

	entries	sub-entries (multiword units)
# entries	35,697	3,230
satisfactory links	80%	33%
links between roots	17%	-

Table 2: Links between the lexical database (EDBL) and the dictionary database (DKB)

These links have been established automatically. Data correspond to the entries of the dictionary that have been linked to EDBL´s entries. The low integration-level of the sub-entries is because they are multiword units and in the lexical database their components are stored separately. In the case of derived forms (Basque is agglutinative), we decided to link the roots when it is not possible to link the whole forms.

With respect to the lexical-semantic part of the DKB, the acquisition of relations from the dictionary is in progress.

Table 3 shows the number of relations that have been extracted from the dictionary and stored in the knowledge base so far.

	Extracted relations	**Stored relations**	
Synonymy	19,809	16,949	%85.6
Hypernymy	20,658	18,331	%88.7
Specific Relators	5,386	4,169	%77.4
Overall	45,853	39,449	%86

Table 3: State of the DKB

About 40,000 relations have been stored. The difference between the number of extracted and stored relations is due, mainly, to circularity problems, that is to say, words that appear in definitions but do not appear as entries. The other important reason is that some relations are duplicated because the morphological analyser yields more than one single analysis for some words. In these cases, we only store one relation and avoid storing the same relation for different analysis.

Finally, the construction of the Basque WordNet knowledge base is in an advanced state (Agirre et al., 2002a). In the last two years, we have been adding Basque senses to EuroWordNet. The current version of Basque WordNet has about 25,500 Basque words and 51,900 senses that have been manually revised.

For the future we are planning to enhance the contents of the lexical-semantic framework. For this purpose we intend to:

- Deal with the relations extracted from a deeper analysis of the dictionary, including the derivational relations.
- Repeat the same process with other bigger monolingual dictionary (Elhuyar, 2000).
- Include relations extracted from other sources, such as corpora, as it is aimed in the MEANING project.
- Incorporate information on named entities and classify them.

6. Conclusions

A general lexical framework has been presented. The representation model is based on assumptions from other models, such as Hiztsua, WordNet, EDR (Artola, 1993; Lenat et al., 1995; Yokoi, 1995) and is adapted to our needs according to the following features:

- Levels of lexical knowledge and connectivity: the morphosyntactic information and the semantic information have been stored in two different databases. Both databases have been linked by means of the lexical unit identifiers.
- In the semantic knowledge base, the distinction between phrasal and type concepts involves two representation layers. For representing conceptual knowledge, a relational model has been adopted.
- This lexical framework is conceived as a general working environment. Special attention has been paid to the state of the stored information, that could be fully or partially desambiguated.

- The representation schema is open, flexible and general. It must be underlined that the extended entity-relation schema of the proposed model subsumes WordNet and EuroWN, even the forthcoming 2.0 version, and allows to represent all the information extracted from the monolingual dictionary. The model could be reused to represent heterogeneous lexical resources.
- The representation of the Basque WordNet includes entities and types of relations that are not present in WordNet.
- The Basque WordNet and the concepts extracted from the monolingual dictionary are being mapped. Our purpose is to enrich the Basque WordNet resource with the relations stored in the Concept Classification level, and vice versa. Indirectly, we would enrich EuroWordNet, given that it is a multilingual resource in which the relations for each language could be used in the others.

7. Acknowledgments

This work was partially funded by the MCYT HERMES project (TIC-2000-0335) and the EC MEANING project (IST-2001-34460).

8. References

(Agirre et al., 1994) Agirre E., Arregi X., Artola X., Díaz de Ilarraza A., Sarasola, K."*Lexical Knowledge Representation in an Intelligent Dictionary Help System*", Proceedings of COLING'94, vol. 1, 544-550. Kyoto (Japan). August 1994.

(Agirre et al., 2000) Agirre E. Ansa O., Arregi X., Artola X., Díaz De Ilarraza A., Lersundi M., Martínez D., Urizar R., Sarasola K. "*Extraction Of Semantic Relations From A Basque Monolingual Dictionary Using Constraint Grammar*" Proceedings of Euralex Sttutgart (Germany). 2000.

(Agirre & Lersundi, 2001) Agirre E., Lersundi M. "*Extracción de relaciones léxico-semánticas a partir de palabras derivadas usando patrones de definición*" Proceedings of SEPLN 2001. Jaén (Spain). September 13-14-15, 2001.

(Agirre et al., 2001) Agirre E., Ansa O., Martínez D. and Hovy E. "*Enriching WordNet concepts with topic signatures*". Proceedings of the NAACL workshop on WordNet and Otherlexical Resources: Applications, Extensions and Customizations. Pittsburg.2001.

(Agirre & Martinez, 2001) Agirre E. and Martinez D. "*Learning class-to-class selectional preferences*". Proceedings of the Workshop "Computational Natural Language Learning"(CoNLL-2001). In conjunction with ACL'2001/EACL'2001. Toulouse, France.2001.

(Agirre et al., 2002a) Agirre E., Ansa O., Arregi X., Arriola J.M., Diaz de Ilarraza A., Pociello E., Uria L. "Methodological issues in the building of the Basque WordNet: quantitative and qualitative analysis", Proceedings of First International WordNet Conference. Mysore (India), 2002

(Agirre et al., 2002b) Agirre E., Lersundi M., Martínez D., "*A Multilingual Approach to Disambiguate Prepositions and Case Suffixes*" ACL Workshop: Word Sense Disambiguation: recent successes and future directions, 2002

(Aldezabal et al., 2001) Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza N., Hernández G., Lersundi M. "*EDBL: a General Lexical Basis for the Automatic Processing of Basque*" IRCS workshop on linguistic databases, Philadelphia, USA. December 11-13, 2001

(Artola X., 1993) Artola X., "*HIZTSUA: Hiztegi-sistema urgazle adimendunaren sorkuntza eta eraikuntza/Conception et construction d'un système intelligent d'aide dictionnariale (SIAD)"*, Donostia 1993, UPV-EHU, PhD Thesis.

(Elhuyar, 2000) Elhuyar, "*Euskal Hiztegi Modernoa*", Elhuyar K.E. Usurbil. 2000.

(Fellbaum, 1998) Fellbaum C. "*WordNet: An Electronic Lexical Database*". The MIT Press, Massachusetts. London, England, 1998.

(Lenat et al., 1995) Lenat D., Miller G., Yokoi T. "*CYC, WordNet, and EDR: Critiques and Responses*", Communications of the ACM, vol. 38, no. 11, 45-48, 1995.

(Quillian, 1968) Quillian M.R.. Semantic Memory, in M. Minsky ed., 227-270, Semantic Information Processing. Cambridge (Mass.): MIT Press, 1968.

(Sarasola I., 1996) Sarasola I., "*Euskal Hiztegia*" Donostia, Gipuzkoako Kutxa, 1996.

(Vossen et al., 1998) Vossen, P., Bloksma L., Climent S., Marti M. A., Oreggioni G., Escudero G., Rigau G., Rodriguez H., Roventini A., Bertagna F., Alonge A., Peters C., Peters W. "*The Reestructured Core wordnets in EuroWordnet: Subset1*". EuroWordNet(LE-4003) Deliverable D014/D015, University of Amsterdam. 1998.

(Vossen et al., 2001) Vossen P., Bloksma L., Climent S., Marti M.A., Taule M., Gonzalo J., Chugur I., Verdejo M.F., Escudero G., Rigau G., Rodriguez H., Alonge A., Bertagna F., Marinelli R., Roventini A., Tarasi L., W. Peters. "*Final Wordnets for Dutch, Spanish, Italian and English*", EuroWordNet (LE2-4003) Deliverable D032/D033, University of Amsterdam. 2001.

(Yokoi, 1995) Yokoi T., "*The EDR electronic Dictionary*", CAMC 38(11), 1995.